UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Milena Marinho Arruda

Contribuições no Contexto da Teoria da Informação para o Processamento de Sinal Genômico

Campina Grande - PB
2022

Milena Marinho Arruda

Contribuições no Contexto da Teoria da Informação para o Processamento de Sinal Genômico

Tese de doutorado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande, pertencente à linha de pesquisa Eletrônica e Telecomunicações e a área de concentração Processamento da Informação, como requisito para obtenção do Título de Doutora em Engenharia Elétrica.

Orientador: Prof. Dr. Francisco Marcos de Assis

Campina Grande - PB
2022

A779c Arruda, Milena Marinho.
Contribuições no contexto da teoria da Informação para o processamento de sinal genômico / Milena Marinho Arruda. – Campina Grande, 2022.
132 f. : il.

Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2022.
"Orientação: Prof. Dr. Francisco Marcos de Assis".
Referências.

1. Eletrônica e Telecomunicações. 2. Códigos BCH. 3. Códigos Corretores de Erros. 4. Processamento de Sinal Genômico. 5. Sequências de DNA. 6. Teoria da Informação e Codificação. 7. Sequências de Codificação. 8. Processamento da Informação. I. Assis, Francisco Marcos de. II. Título.

CDU 621.391(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

Contribuições no Contexto da Teoria da Informação para o Processamento de Sinal Genômico

MILENA MARINHO ARRUDA

TESE APROVADA EM 07/10/2022

FRANCISCO MARCOS DE ASSIS, Dr., UFCG
Orientador(a)

BENEMAR ALENCAR DE SOUZA, D.Sc. , UFCG
Examinador(a)

HELDER ALVES PEREIRA, Dr., UFCG
Examinador(a)

DANILO SILVA, Ph.D., UFSC
Examinador(a)

GIULIANO GADIOLI LA GUARDIA, Dr., UEPG-PR
Examinador(a)

CHARLES CASIMIRO CAVALCANTE, Dr., UFC
Examinador(a)

CAMPINA GRANDE - PB

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM ENGENHARIA ELETRICA
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

REGISTRO DE PRESENÇA E ASSINATURAS

1. ATA DA DEFESA PARA CONCESSÃO DO GRAU DE DOUTOR EM CIÊNCIAS, NO DOMÍNIO DA ENGENHARIA ELÉTRICA, REALIZADA EM 07 DE OUTUBRO DE 2022 **(Nº 353)**

CANDIDATA: **MILENA MARINHO ARRUDA**. COMISSÃO EXAMINADORA: BENEMAR ALENCAR DE SOUZA, D.Sc. , UFCG, Presidente da Comissão e Examinador Interno, FRANCISCO MARCOS DE ASSIS, Dr., UFCG, Orientador, CHARLES CASIMIRO CAVALCANTE, Dr., UFC, GIULIANO GADIOLI LA GUARDIA, Dr., UEPG-PR, DANILO SILVA, Ph.D., UFSC, Examinadores externos. TÍTULO DA TESE: Contribuições no Contexto da Teoria da Informação para o Processamento de Sinal Genômico. ÁREA DE CONCENTRAÇÃO: Processamento da Informação. HORA DE INÍCIO: **14h00** – LOCAL: **Sala Virtual, conforme Art. 5° da PORTARIA SEI Nº 01/PRPG/UFCG/GPR, DE 09 DE MAIO DE 2022**. Em sessão pública, após exposição de cerca de 45 minutos, o(a) candidato(a) foi arguido(a) oralmente pelos membros da Comissão Examinadora, tendo demonstrado suficiência de conhecimento e capacidade de sistematização, no tema de sua tese, obtendo conceito APROVADO. Face à aprovação, declara o(a) presidente da Comissão, achar-se o examinado, legalmente habilitado(a) a receber o Grau de Doutor em Ciências, no domínio da Engenharia Elétrica, cabendo a Universidade Federal de Campina Grande, como de direito, providenciar a expedição do Diploma, a que o(a) mesmo(a) faz jus. Na forma regulamentar, foi lavrada a presente ata, que é assinada por mim, Filipe Emmanuel Porfírio Correia, e os membros presentes da Comissão Examinadora. Campina Grande, 7 de Outubro de 2022.

FILIPE EMMANUEL PORFÍRIO CORREIA

Secretário

BENEMAR ALENCAR DE SOUZA, D.Sc. , UFCG

Presidente da Comissão e Examinador Interno

FRANCISCO MARCOS DE ASSIS, Dr., UFCG

Orientador

CHARLES CASIMIRO CAVALCANTE, Dr., UFC

Examinador Externo

GIULIANO GADIOLI LA GUARDIA, Dr., UEPG-PR

Examinador Externo

DANILO SILVA, Ph.D., UFSC

Examinador Externo

MILENA MARINHO ARRUDA

Candidata

2 - APROVAÇÃO

2.1. Segue a presente Ata de Defesa de Tese de Doutorado da candidata MILENA MARINHO ARRUDA, assinada eletronicamente pela Comissão Examinadora acima identificada.

2.2. No caso de examinadores externos que não possuam credenciamento de usuário externo ativo no SEI, para igual assinatura eletrônica, os examinadores internos signatários **certificam** que os examinadores externos acima identificados participaram da defesa da tese e tomaram conhecimento do teor deste documento.

Documento assinado eletronicamente por **FILIPE EMMANUEL PORFIRIO CORREIA**, **ASSISTENTE EM ADMINISTRACAO**, em 14/10/2022, às 16:34, conforme horário oficial de Brasília, com fundamento no art. 8°, caput, da Portaria SEI n° 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **FRANCISCO MARCOS DE ASSIS**, **PROFESSOR 3 GRAU**, em 17/10/2022, às 08:59, conforme horário oficial de Brasília, com fundamento no art. 8°, caput, da Portaria SEI n° 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **BENEMAR ALENCAR DE SOUZA**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 20/10/2022, às 09:20, conforme horário oficial de Brasília, com fundamento no art. 8°, caput, da Portaria SEI n° 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **Milena Marinho Arruda**, **Usuário Externo**, em 31/10/2022, às 17:29, conforme horário oficial de Brasília, com fundamento no art. 8°, caput, da Portaria SEI n° 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **2832985** e o código CRC **EEBA549E**.

À minha irmã Aline Marinho Arruda.

Agradecimentos

Agradeço, primeiramente, aos meus pais, Alex e Gildete, por acreditarem sempre em mim e nas minhas responsabilidades perante aos compromissos da vida, e, sobretudo, obrigada pela lição de amor que me ensinaram. À minha irmã, Aline, afinal, ter uma irmã é ter, pra sempre, uma vida lembrada com segurança em outro coração. Às minhas primas e irmãs, Beatriz e Gabriela, que a vida me possibilitou escolher, e me suportaram nos momentos mais inusitados.

À minha família e amigos que vibraram minhas conquistas, carregando a certeza de que nunca estarei só. Vocês foram essenciais para que meus dias se tornassem mais leves e divertidos. A todas as pessoas que muitas vezes, anonimamente, fizeram uma diferença enorme na minha vida.

Aos voluntários e voluntárias do Ramo Estudantil IEEE UFCG, em especial do IEEE Women in Engineering UFCG, com os quais descobri que é possível desenvolver habilidades de liderança, comunicação e trabalho em grupo, mesmo em uma rotina diferenciada quanto esta que enfrentamos no mestrado acadêmico. Com eles também consegui formar uma rede de frIEEEnds, ao redor do Brasil e do mundo, com os quais aprendi a importância de devolver à sociedade todo o conhecimento que adquirimos ao longo de nossa formação.

À todos os Professores, mestres e doutores que foram responsáveis por minha formação, dedicando seu apoio, atenção, paciência, amizade e compreensão. Em especial, agradeço ao meu orientador, Professor Francisco Marcos, pela confiança, por estar sempre à disposição moldando o meu conhecimento e, além de tudo, pelos conselhos sobre a vida pessoal e profissional que eu levarei para sempre. À Professora e amiga Luciana Veloso por todas as conversas que incluíam não apenas torca de conhecimento técnico, mas também grandes lições de vida.

Agradeço aos colegas de trabalho do IQuanta, sejam eles professores, alunos, e ex-alunos, dos quais tive apoio e com quem tive oportunidade de conviver e dividir conhecimento. Em especial à Andresso, Juliana, Micael, Taciana e Thiciany, pelas conversas descontraídas e debates científicos.

Ao Programa de Pós-Graduação em Engenharia Elétrica (PPgEE - COPELE) da UFCG, pelo suporte administrativo. Ao CNPq, pelo suporte financeiro para o desenvolvimento dessa tese.

"Talvez um dia, para além dos dias,
Encontres o que queres porque o queres."
Fernando Pessoa

Resumo

O crescimento dos bancos de dados biológicos e a necessidade de compreender como os muitos componentes presentes em uma célula viva estão interagindo e trabalhando juntos para execução de funções celulares são razões que justificam a aplicação interdisciplinar de teorias matemáticas, estatísticas e computacionais para análise e processamento da informação genômica. A informação genética de um organismo está codificada em moléculas de ácido desoxirribonucleico (DNA, do inglês: *deoxyribonucleic acid*) por meio de unidades denominadas bases. A análise e o processamento de sequências de DNA para obtenção de conhecimento biológico constituem o domínio deste documento de tese. A pesquisa desenvolvida visa integrar a teoria e os métodos de processamento de sinais e a teoria da informação para extração de informações genômicas. Um dos principais desafios é, portanto, definir uma regra de mapeamento para representação de sequências de DNA que estão, inicialmente, em um domínio simbólico, e levá-las para um domínio numérico. O primeiro resultado apresentado nesta tese considera um mapeamento unidimensional bijetivo para elementos de um corpo finito com o objetivo de analisar a hipótese de que o DNA está atuando como um código linear na transmissão da informação armazenada. Dessa maneira, existiria um código de correção de erros subjacente às sequências de DNA. Nesse contexto, é proposto um novo algoritmo para buscar códigos BCH cujas palavras-código estão a uma distância de Hamming no máximo unitária do vetor numérico resultante do mapeamento de uma dada sequência de DNA. Além disso, é demonstrado que as sequências de DNA estão distribuídas de maneira aproximadamente uniforme, sob a métrica de Hamming, em um espaço vetorial de dimensão n. Sendo assim, os polinômios geradores dos códigos que identificam coleções de sequências taxonomicamente próximas não fornecem informações biológicas suficientes para agrupar e classificar tais coleções. O segundo resultado apresentado foi alcançado com base na hipótese de que ao considerar um mapeamento fixo para todas as sequências de DNA não é possível garantir que as características intrínsecas de cada sequência estarão sendo devidamente extraídas. Portanto, são propostos dois novos algoritmos: SNR-SE e TBP-SE, ambos baseados na teoria de envoltória espectral para o cálculo desses mapeamentos. A aplicabilidade desses métodos no contexto da análise espectral para discriminação de sequências codificantes e não codificantes de proteínas é analisada e comparada com outros mapeamentos já consolidados na literatura. Nesse cenário, o algoritmo proposto, TBP-SE, teve a maior

acurácia e sensibilidade entre todos avaliados. Destacando-se assim, uma vez que, nesta aplicação a sensibilidade é especialmente importante, pois, assim, a probabilidade de ter uma sequência de codificação que não será identificada é baixa. Além disso, o TBP-SE demonstrou bom desempenho até mesmo para detectar regiões com sequências de codificação mais curtas.

Palavras-chave: Códigos BCH. Códigos Corretores de erros. Processamento de Sinal Genômico. Sequências de DNA. Sequências de codificação. Teoria da Informação e Codificação.

Abstract

The growth of biological databases and the need to understand how the many components present in a living cell are interacting and working together to perform cellular functions are reasons that justify the interdisciplinary application of mathematical, statistical and computational theories for the analysis and processing of genomic information. The genetic information of an organism is encoded in deoxyribonucleic acid molecules (DNA) by means of units called bases. The analysis and processing of DNA sequences to obtain biological knowledge constitute the domain of this document. The research developed aims to integrate the theory and methods of signal processing and information theory to extract genomic information. One of the main challenges is, therefore, to define a mapping rule to represent DNA sequences that are initially in a symbolic domain, taking them to a numerical domain. The first result considers a bijective unidimensional mapping for elements of a finite field with the aim of analyzing the hypothesis that DNA is acting as a linear code in the transmission of stored information. Hence, there will be an error-correcting code underlying the DNA sequences. In this context, a new algorithm is proposed to search for BCH codes whose codewords are at a Hamming distance at most unity from the numerical vector resulting from the mapping of a given DNA sequence. Furthermore, it is shown that the DNA sequences are approximately uniformly distributed, under the Hamming metric, in a vector space of dimension n. Therefore, the genrator polynomial of the codes that identify collections of taxonomically close sequences do not provide enough biological information to group and classify them. The second result based on the hypothesis that when considering a fixed mapping for all DNA sequences, it is not possible to guarantee that the intrinsic characteristics of each sequence will be properly extracted. Therefore, two new algorithms are proposed: SNR-SE and TBP-SE, both based on the spectral envelope theory to calculate these mappings. The applicability of these methods in the context of spectral analysis to discriminate coding and non-coding sequences of proteins is analyzed and compared with other mappings already consolidated in the literature. In this scenario, the proposed algorithm, TBP-SE, had the highest accuracy and sensitivity among all evaluated. This stands out, since, in this application, sensitivity is especially important, as the probability of having a coding sequence that will not be identified is low. In addition, TBP-SE demonstrated good assertiveness even to detect regions with shorter

coding sequences.

Keywords: BCH Codes. DNA sequences. Error Correcting Codes. Genomic Signal Processing. Coding sequence. Information Theory.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

BCH	Código Bose-Chaudhuri-Hocquenghem
CDS	*Coding Sequence*
CGR	*Chaos Game Representation*
DFT	Transformada de Fourier em Tempo Discreto
DNA	*Deoxyribonucleic acid*
EIIP	Potencial de Interação Elétron-Íon
FFT	Transformada Rápida de Fourier
FN	Falso Negativo
FP	Falso Positivo
FPR	Taxa de Falso Positivo
MEM	*Minimum Entropy Mapping*
NCBI	*National Center for Biotechnology Information*
PAM	*Pulse Amplitude Modulation*
PGZ	Decodificador Peterson–Gorenstein–Zierler
QPSK	*Quadrature Phase Shift Keying*
RNA	*Ribonucleic acid*
SE	Envoltória Espectral
SNP	*Single Nucleotide Polymorfism*
SNR	Relação Sinal Ruído
STFT	Transformada de Fourier de Tempo Reduzido

TBP	*Three-Base Periodicity*
TN	Verdadeiro Negativo
TNR	Taxa de Verdadeiro Negativo
TP	Verdadeiro Positivo
TPR	Taxa de Verdadeiro Positivo

Lista de Símbolos

$\mathbb{F}_q[x]$	Anel de polinômio na variável x
$\mathcal{C}$	Código linear
$\mathcal{C}_{BCH}$	Código BCH
$\mathbb{F}_q$	Corpo finito de ordem q
$H(\cdot)$	Entropia
$S[k]$	Espectro de energia
$\mathcal{M}$	Mapeamento
mmc	Mínimo múltiplo comum
$[n, k, d]$	Parâmetros de um código linear
$g(x)$	Polinômio gerador de um código BCH
$x[n]$	Sinal discreto no tempo
s_{n-k+1}^{n}	Subsequência de comprimento k: $x_{n-k+1}x_{n-k+2}\cdots x_n$
s	Sequência de DNA de comprimento N
$\mathbf{x}$	Vetor n-dimensional

Sumário

Parte I

Introdução

Capítulo 1

Introdução

Neste capítulo serão apresentados a motivação, as contribuições e a organização do texto deste documento. No cenário da motivação inclui-se a revisão bibliográfica, e o estado da arte. Além disso, as contribuições destacam os pontos nos quais esta pesquisa colaborou com a comunidade científica.

1.1 Motivação

A combinação das áreas de biologia, ciência da computação, estatística, matemática e engenharia resulta na grande área de pesquisa da biologia computacional, uma área interdisciplinar que corresponde à aplicação dessas ciências para análise e processamento da informação nas áreas de estudo da biologia. Uma das abordagens para a análise de dados genômicos que tem atraído a atenção da comunidade científica nos últimos anos corresponde ao processamento de sinais genômicos (GSP, do inglês: *genomic signal processing*). O crescimento dos bancos de dados biológicos e a necessidade de compreender como os muitos componentes presentes em uma célula viva estão interagindo e trabalhando juntos para execução de funções celulares são razões que justificam o interesse em ferramentas matemáticas, estatísticas e computacionais [1].

O ponto de partida é que a célula viva é um sistema no qual muitos componentes estão interagindo e trabalhando juntos para execução de funções celulares e interação com o meio. Para entender o seu comportamento, um modelo desses componentes e suas interações é exigido. Assim sendo, o GSP se refere ao uso da Teoria de Processamento de Sinais Digitais, incluindo reconhecimento de padrões, Teoria da Informação, Sistemas Dinâmicos, Teoria de Controle e Teoria da Comunicação, para a análise das informações genômicas.

A informação genética de um organismo é codificada em moléculas de ácido desoxirribonucleico (DNA, do inglês: *deoxyribonucleic acid*) por meio de unidades chamadas bases: adenina (`A`), citosina (`C`), guanina (`G`) e timina (`T`). As análises no contexto do GSP, são feitas, portanto, mapeando uma determinada sequência de DNA,

cujo domínio é simbólico, em um sinal numérico [2, 3]. Ao mapear adequadamente uma sequência de caracteres em um ou mais sinais discretos, as diversas técnicas de processamento de informação podem ser aplicadas. A transformada de Fourier, por exemplo, ao ser adequadamente definida, pode ser usada para extrair informações importantes de periodicidades das bases do DNA. Porém, existe um desafio em definir uma regra de mapeamento ideal para tais sequências.

A maioria dos métodos de GSP são aplicados para identificação de sequências codificantes de proteínas em sequências de DNA [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. Contudo, outras aplicações incluem a determinação das propriedades estruturais, e termodinâmicas do DNA [15], busca de repetições genômicas [16, 17], estimativa de similaridade e alinhamento de sequências de DNA [18, 19, 20, 21, 22] e classificação filogenética [23, 24, 25, 26, 27, 28]. Embora as aplicações dos métodos de GSP sejam diversas, limita-se o escopo desta tese. Portanto, segue o estado da arte bem como, a formulação dos problemas a serem investigados em cada contexto.

Teoria dos Códigos

No contexto da aplicação da Teoria dos Códigos, as abordagens da Teoria da Informação e Codificação têm sido investigadas em áreas da bioinformática, por exemplo uso do DNA como meio de armazenar e ocultar informação [29, 30, 31, 32], além de buscar códigos corretores de erros subjacentes às sequências de DNA [33, 34]. Essa última aplicação é alvo de investigação deste documento de tese. Pesquisadores tentam identificar similaridades entre os sistemas biológicos e de comunicação para responder questões como: existe um mecanismo de controle de erro nas sequências biológicas similar aos códigos corretores de erros empregados em sistemas digitais? [35, 36, 37, 38, 39]; existem códigos capazes de identificar e/ou reproduzir tais sequências? [40, 41, 42, 43, 44].

Embora relevantes, tais perguntas ainda não têm uma resposta ou modelo definitivo. Liebovitch *et al.* [40] foram os primeiros a introduzir uma metodologia para determinar se um código linear de correção de erro está presente nas sequências de DNA. Rosen [41] continuou esta investigação apresentando um método para descobrir uma estrutura de código de correção de erros em sequências para, então, detectar repetições tandem (repetições que ocorrem no DNA quando um padrão de uma ou mais bases é repetido de forma adjacente). Faria *et al.* [42] e Rocha *et al.* [43] foram mais específicos e propuseram um algoritmo, conhecido como *DNA Sequence Generator*, que verifica se uma determinada sequência de DNA pode ser identificada como palavras-código de um código BCH de distância de projeto $d = 3$ sobre corpos finitos e anéis de inteiros, respectivamente. Nesse contexto, diz-se que uma sequência de DNA é identificada por um código BCH se tal sequência pode ser mapeada para um vetor que é palavra-código do código ou difere de alguma palavra-código em até um símbolo. No contexto biológico, essa incompatibilidade é conhecida como polimorfismo de nucleotídeo único (SNP, do inglês: *Single Nucleotide*

Polymorphism).

Algumas das aplicações do algoritmo *DNA Sequence Generator* encontradas na literatura são citadas a seguir. Faria *et al.* [44] usou o algoritmo para mostrar que um gene e até mesmo um genoma plasmidial podem ser identificados como palavras-código de códigos BCH. Brandão *et al.* [45] utilizou uma ferramenta para avaliar o caminho evolutivo do código genético de alguns genes e investigar o significado biológico do descasamento único entre a sequência de DNA original e a sequência identificada como a palavra-código. Duarte-González *et al.* [39] usou o algoritmo para representar sequências de proteínas e explorar relações evolutivas. A fim de resolver algumas restrições em relação ao comprimento da sequência de DNA e reduzir o esforço computacional deste algoritmo, Rodríguez-Sarmiento *et al.* [33] e Hernádez *et al.* [34] propuseram um novo algoritmo, baseado na técnica de fatoração polinomial, para identificar sequências biológicas de comprimento ímpar usando códigos BCH sobre corpos finitos e anéis de inteiros, respectivamente.

Diante das limitações no processo de decodificação dos algoritmos para identificação de sequências de DNA usando códigos BCH, propõe-se, portanto, explorar o decodificador Peterson–Gorenstein–Zierler (PGZ) em tais aplicações. A partir dessa abordagem propõe-se um novo algoritmo para realizar a busca por códigos BCH que identificam sequências de DNA.

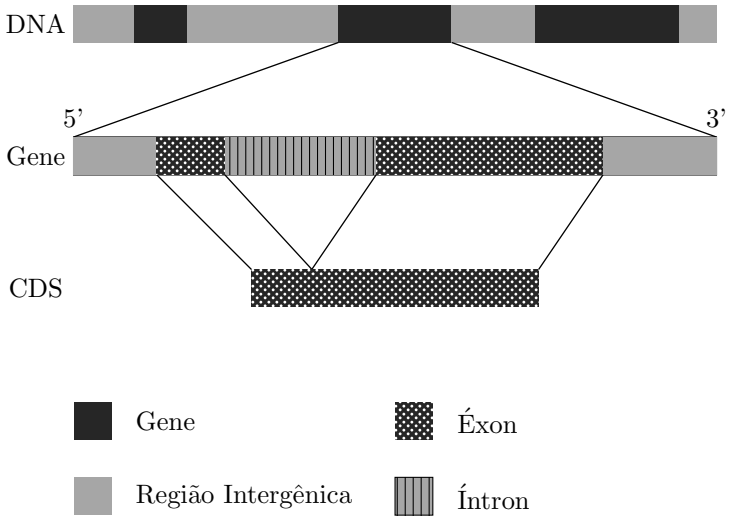
Nesse sentido, surge o seguinte questionamento: uma vez que existe um código de correção de erros subjacente a uma sequência de DNA, esse código revela semelhanças entre sequências de DNA de organismos próximos em uma árvore filogenética? Em caso afirmativo, uma vez que representações numéricas e gráficas de sequências de DNA já foram usadas para propor métodos de classificação sem alinhamento de sequências de DNA [26, 27, 28], os códigos BCH também podem ser usados como um método de classificação sem alinhamento? Para responder a essas questões, propõe-se analisar a significância estatística de encontrar tais códigos. Para tanto, compara-se a probabilidade de um código BCH identificar um vetor aleatório com a probabilidade de identificar uma sequência de DNA verdadeira.

Análise Espectral

A análise espectral de sequências de DNA tem sido amplamente investigada como um indicador para discriminação de sequências codificantes (CDS, do inglês: *Coding Sequence*) e não codificantes de proteína do DNA. Nesse contexto, Trifonov [46] observou a existência de periodicidades em sequências de DNA a partir da análise da função de autocorrelação; e Fickett [47] observou que enquanto as regiões não codificantes mostram um padrão bastante aleatório, as sequências codificantes revelam periodicidades; em particular, a periodicidade de três bases (TBP, do inglês: *Three-Base Periodicity*).

Nas células eucarióticas, o DNA é dividido em regiões gênicas e intergênicas. Os genes

Figura 1.1 – Representação das diferentes regiões do DNA eucarioto.



Fonte: Elaborada pela autora.

são divididos em éxons e íntrons. As sequências codificadoras de proteínas são então a porção de um gene que codifica uma proteína: seus éxons. A região codificadora de um gene também é conhecida como sequência de codificação. As sequências não codificantes referem-se aos íntrons e regiões intergênicas. Essas regiões são ilustradas na Figura 1.1.

Embora, intuitivamente, a primeira hipótese com relação à característica de periodicidade em sequências codificantes seja devido à natureza tripla do códon, a mesma se mostrou insuficiente. Esse fenômeno periódico tem intrigado muitos biólogos que buscam compreendê-lo e explicá-lo [48, 49, 50, 51, 52]. Shepherd [48] verificou que as formas ancestrais dos genes atuais podem ter consistido na repetição dos códons RNY (purina-qualquer-pirimidina). Tsonis *et al.* atribui essa propriedade ao fato de que alguns aminoácidos são mais predominantes nas proteínas que outros, ou seja, a distribuição dos aminoácidos é enviesada. Howe [52] apesar de concordar que os vieses nos códons e que a distribuição de aminoácidos contribuam para essa periodicidade, considera que a verdadeira fonte da periodicidade não reside no fato de que os vieses existem, mas de que esses vieses observados levam a distribuições multinomiais desiguais de nucleotídeos em função das suas posições nos códons. Por essa razão, a análise de Fourier é útil, um vez que é capaz de revelar periodicidades a partir da observação do espectro de energia de um sinal.

Um método clássico para análise espectral foi proposto por Voss [53], no qual cada uma das quatro bases está associada a um sinal indicador binário. Cada indicador binário é um sinal de tempo discreto que assume 1 quando o n-ésimo símbolo da sequência é uma determinada base e 0 caso contrário. Nesse caso, o espectro de densidade de energia é a soma da contribuição de energia de cada sinal indicador binário avaliado a partir da transformada de Fourier de cada sinal. Além disso, são comuns abordagens nas quais uma sequência de DNA é mapeada para um único sinal numérico. Neste caso, o

espectro de energia é avaliado a partir da DFT desse sinal. Entre os mapeamentos mais comuns, Lalović *et al.* [54] propuseram um mapeamento baseado nos pseudopotenciais de interação elétron-íon (EIIP, do inglês: *electron-ion interaction pseudopotential*); Anastassiou [5] propôs que a imagem do mapeamento é dada por números complexos, semelhante ao esquema de modulação por chaveamento de fase em quadratura (QPSK, do inglês: *Quadrature Phase Shift Keying*); e Galleani *et al.* [8] propuseram o espectro de mapeamento de entropia mínima (MEM, do inglês: *Minimum Entropy Mapping*), no qual um mapeamento adaptativo com imagem real é calculado a partir do critério de minimização de entropia espectral.

No entanto, essas abordagens têm algumas limitações de desempenho, principalmente no que concerne à definição desses mapeamentos. As sequências simbólicas possuem uma estrutura estatística que fornece informações importantes sobre as mesmas. Espera-se, portanto, que a representação numérica de tal sequência não imponha características adicionais ao sinal resultante. Por exemplo, um mapeamento não pode assumir que um símbolo é sempre numericamente maior que outro. Por essa razão, fica claro que o mesmo mapeamento para qualquer sequência de DNA deve ignorar as características que lhe são particularmente inerentes. Assim, isso sugere que para cada sequência de DNA, um mapeamento específico deve ser realizado, ou seja, deve-se definir mapeamentos adaptativos para tais sequências.

Assumindo que um sinal numérico é apropriado para uma determinada sequência de DNA, então a análise espectral pode ser aplicada para detectar regiões codificantes em genes. Nesse sentido, Tiwari *et al.* [4] foram os primeiros pesquisadores a propor que é suficiente avaliar a densidade de energia na frequência $1/3$ rad/amostra em uma janela de amostras, deslizando-a por toda a sequência de DNA. Vaidyanathan *et al.* [55] propuseram o uso do filtro *antinotch* na janela deslizante. Sahu *et al.* [9] sugeriram o uso da transformada S e Roy *et al.* [10] o uso de um estimador de norma mínima, ambos considerando o sinal resultante do mapeamento EIIP. Wang e Johnson [56] expandiram a abordagem da envoltória espectral (inicialmente proposta por Stoffer *et al.* [57]) para processar sinais simbólicos não estacionários no domínio tempo-frequência e analisaram a estrutura de correlação do DNA.

Nesse contexto, investigou-se a problemática do mapeamento adaptativo, em que cada sequência de DNA deve ser mapeada para um sinal numérico conforme algum critério preestabelecido. Portanto, neste documento propõe-se dois novos algoritmos para calcular mapeamentos adaptativos para sequências de DNA. Esses mapeamentos são, então, aplicados para a análise espectral das respectivas sequências com o objetivo de melhorar a discriminação entre sequências codificantes e não codificantes. O primeiro algoritmo proposto procura o mapeamento que maximiza a SNR do espectro de energia. O segundo algoritmo, por outro lado, aproveita o conhecimento prévio sobre a propriedade TBP. Nesse caso, o mapeamento resulta da envoltória espectral na frequência $k = \lfloor N/3 \rfloor$

em que N é o comprimento da sequência. Além disso, o desempenho dos novos métodos é verificado comparando-os com o desempenho de outros quatro métodos bem estabelecidos na literatura — Voss [53], EIIP [58], QPSK [5] e MEM [8] — e aplicando-os à sequências de DNA reais e simuladas cujas propriedades são conhecidas.

1.2 Base de Dados

A capacidade de sequenciar proteínas, ácidos nucleicos e outras sequências de polímeros de um organismo tornou-se uma das ferramentas mais importantes na pesquisa biológica moderna [59]. Assim sendo, foram desenvolvidos diversos bancos de dados públicos que coletam, verificam e publicam sequências de todo o mundo, entre os quais a Colaboração Internacional de Banco de Dados de Sequência de Nucleotídeos (INSDC, do inglês: *International Nucleotide Sequence Database Collaboration*) que é uma iniciativa que opera entre:

1. Banco de Dados de DNA do Japão (DDBJ, do inglês *DNA Data Bank of Japan*): <https://www.ddbj.nig.ac.jp/>;

2. Instituto Europeu de Bioinformática (EMBL-IBI, do inglês: *European Molecular Biology Laboratory - European Bioinformatics Institute*): <https://www.ebi.ac.uk/ena>;

3. Centro Nacional de Informação Biotecnológica (NCBI, do inglês: *National Center for Biotechnology Information*): <https://www.ncbi.nlm.nih.gov/>.

Nesse documento de Tese, todos os resultados envolvendo sequências de DNA são de sequências que estão disponíveis no banco de dados de nucleotídeos do NCBI, uma vez que este fornece acesso aberto a informações biomédicas e genômicas [60]. As sequências têm um identificador, o número GI, que é uma série simples de dígitos processados pelo NCBI que a identificam. Portanto, ao longo do texto uma sequência de DNA é, também, referenciada pelo seu número GI.

1.3 Contribuições e Produção Científica

As contribuições desta tese de doutorado que a distingue de outros trabalhos estão listadas abaixo:

- Proposta de um novo algoritmo para buscar códigos BCH que identificam sequências de DNA. Nesse caso, busca-se por códigos BCH cujas palavras-códigos diferem em até um símbolo do vetor, devidamente mapeado, de uma sequência de DNA;

- Avaliação de como as sequências de DNA de uma mesma classificação taxonômica, quando devidamente representadas por vetores numéricos, estão distribuídas em um espaço vetorial. Dessa forma, avaliou-se a possibilidade de caracterização de conjuntos de sequências utilizando-se códigos BCH;
- Proposta de dois novos métodos baseados na envoltória espectral para determinação de mapeamentos adaptativos para sequências de DNA;
- Aplicação dos mapeamentos adaptativos propostos para classificação e discriminação de sequências codificantes a partir da análise espectral;
- Análise comparativa dos algoritmos e métodos propostos com os existentes na literatura.

Os resultados dessas contribuições foram publicados em congresso nacional e periódicos da área, sendo eles:

- ARRUDA, M. M.; SILVA, A. da; ASSIS, F. M. An Adaptive Mapping Method Using Spectral Envelope Approach for DNA Spectral Analysis. Entropy, v. 24, n. 7, 2022. ISSN 1099-430. [61]
- ARRUDA, M. M.; SILVA, A. da; ASSIS, F. M. Maximizing the SNR of DNA Spectrum for Coding Sequence Identification. In: Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais. Sociedade Brasileira de Telecomunicações, 2021. [62]
- SILVA, A. da; ARRUDA, M. M.; ASSIS, F. M. Reconstrução de Árvores Filogenéticas a partir de mtDNA usando o Algoritmo SEQUITUR. In: Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais. Sociedade Brasileira de Telecomunicações, 2021. [63]
- ARRUDA, M. M.; ASSIS, F. M.; SOUZA, T. A. Is BCH Code Useful to DNA Classification as an Alignment-Free Method? IEEE Access, v. 9, p. 68552–68560, 2021. ISSN 2169-3536. [64]

1.4 Organização do Documento

Este documento está organizado em seis capítulos. Este capítulo de introdução apresentou a motivação e os objetivos deste documento de tese. No Capítulo 2 será apresentada uma breve revisão das técnicas de processamento de sinais, computação e Teoria da Informação usadas para alcançar os objetivos nas respectivas áreas de aplicação. O objetivo do Capítulo 3 é integrar a teoria e os métodos para extração de informação genômica. No Capítulo 4 abordaremos uma problemática específica da Teoria

de Códigos aplicada à bioinformática: identificação do DNA a partir de códigos BCH. Serão apresentadas melhorias para um algoritmo já existente na literatura. Além disso, propomos um novo algoritmo e analisamos a sua performance computacional e estatística. O Capítulo 5 abordará a problemática de discriminação de sequências codificantes a partir da análise espectral de tais sequências. Serão propostos dois algoritmos para determinar mapeamentos adaptativos para sequências de DNA, e assim, melhorar o uso da análise espectral como critério de discriminação de sequências codificantes de proteína. As considerações finais ao que concerne cada capítulo de resultados serão apresentadas nos respectivos capítulos. Por fim, no Capítulo 6 serão apresentadas as conclusões gerais e as sugestões de trabalhos futuros. Além dos capítulos citados, o presente trabalho apresenta no Apêndice A os conceitos básicos da biologia molecular que são importantes para acompanhar e complementar o contexto biológico tratado ao longo deste documento. No Apêndice B é apresentado uma breve descrição das estruturas algébricas e definições de códigos corretores de erros relevantes para o entendimento dos códigos BCH. No Apêndice C é apresentado a prova do quociente de Rayleight, cujo resultado foi utilizado no Capítulo 5. Nos Apêndices D e E são apresentados os detalhes para o cálculo da distância de Levenshtein e do algoritmo de agrupamento UPGMA, respectivamente. Por fim, no Apêndice F são apresentadas as representações gráficas por meio do CGR do DNA de alguns mamíferos.

Parte II

Fundamentos

Capítulo 2

Fundamentos Teóricos

Neste capítulo, uma breve revisão das técnicas de processamento de sinais, computação e Teoria da Informação é apresentada. Esse fundamentos e as formulações matemáticas envolvidas serão úteis para o progresso dos problemas que esta tese se propõe a investigar.

2.1 Teoria da Informação

A Teoria da Informação é uma ciência originalmente proposta em 1948 por Claude Shannon por meio do artigo *A Mathematical Theory of Communication* [65]. Neste artigo são definidas importantes medidas de informação com a finalidade de quantificar os limites fundamentais nas operações de processamento de sinais e comunicação confiável de dados.

2.1.1 Medidas de Informação

As medidas de informação são úteis para quantificar a dependência ou causalidade entre variáveis ou processos aleatórios. Dois eventos são independentes quando a ocorrência de um não é influenciada pela ocorrência do outro. Da mesma forma, duas variáveis aleatórias são independentes se a realização de uma não afeta a distribuição de probabilidade da outra. Ao considerar dois eventos x e y, associados às variáveis aleatórias X e Y, respectivamente, essas variáveis aleatórias são ditas independentes se, e somente se, $\mathrm{P}(x, y) = \mathrm{P}(x)\mathrm{P}(y)$, para todo x e y, em que P denota a distribuição de probabilidade.

Entropia

O conceito de entropia foi introduzido por Claude Shannon [65] e refere-se à medida de incerteza de uma variável aleatória. Por definição, a entropia de uma variável aleatória discreta X associada a um alfabeto $\mathcal{X}$ é dada por:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} \mathrm{P}(x) \log \mathrm{P}(x). \tag{2.1}$$

Ao longo deste documento, será utilizada a convenção de que $0 \log 0 = 0$, justificada pelo fato de que $\lim_{x \to 0} x \log x = 0$. Além disso, o logaritmo natural é sempre usado; portanto, a entropia é dada em *nats*. Uma consequência imediata da definição é que a entropia discreta é sempre positiva, ou seja, $H(X) \geq 0$.

A definição de entropia pode ser estendida para mais de uma variável aleatória por meio da regra da cadeia, que estabelece a seguinte relação:

$$\mathrm{H}(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} \mathrm{H}(X_i | X_{i-1}, \cdots, X_1). \tag{2.2}$$

Para o caso de um par de variáveis X e Y, a Equação (2.2) se reduz a:

$$\begin{aligned} \mathrm{H}(X, Y) &= \mathrm{H}(X) + \mathrm{H}(Y|X) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{P}(x, y) \log \mathrm{P}(x, y), \end{aligned} \tag{2.3}$$

em que, $\mathrm{H}(X|Y)$ é a entropia condicional, que mede a incerteza da variável aleatória X dado que Y é conhecida.

Divergência de Kullback-Leibler

A divergência de Kullback-Leibler é uma medida entre duas distribuições de probabilidade. Em outras palavras, uma medida de ineficiência (excesso de bits) para codificar uma variável aleatória X, ao assumir uma distribuição $\mathrm{Q}(x)$ quando a distribuição verdadeira é $\mathrm{P}(x)$. A divergência de Kullback-Leibler é, portanto, definida como,

$$D(\mathrm{P}||\mathrm{Q}) = \sum_{x \in \mathcal{X}} \mathrm{P}(x) \log \frac{\mathrm{P}(x)}{\mathrm{Q}(x)}. \tag{2.4}$$

Informação Mútua

A informação mútua foi introduzida por Robert Fano [66] como sendo uma medida de dependência entre variáveis aleatórias. Essa medida quantifica a redução da incerteza de uma variável aleatória dado o conhecimento de outra variável aleatória. A informação mútua entre as variáveis aleatórias discretas X e Y é dada por:

$$\begin{aligned} \mathrm{I}(X; Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \mathrm{P}(x, y) \log \frac{\mathrm{P}(x, y)}{\mathrm{P}(x)\mathrm{P}(y)} \\ &= \mathrm{H}(X) - \mathrm{H}(X|Y). \end{aligned} \tag{2.5}$$

Uma consequência direta da definição é que a informação mútua para quaisquer duas variáveis aleatórias é sempre maior ou igual a zero, com igualdade se, e apenas se, X e Y são independentes.

2.1.2 Códigos BCH

Os códigos BCH formam uma classe de códigos lineares propostos em 1959 por Alexis Hocquenghem [67] e independentemente por Raj Bose e Ray-Chaudhuri [68] em 1960. A construção dos códigos BCH está fundamentada em estruturas algébricas. Recomenda-se, portanto, a leitura do Apêndice B para melhor entendimento dos conceitos discutidos nesta seção.

Considerando um corpo finito $\mathbb{F}_q$ com ordem q, em que q é uma potência de um primo, para qualquer inteiro positivo m, $\mathbb{F}_{q^m}$ é um corpo de extensão de $\mathbb{F}_q$. O espaço vetorial de todos os vetores de comprimento n sobre $\mathbb{F}_q$ é denotado por $\mathbb{F}_q^n$. Se α é um elemento primitivo do corpo $\mathbb{F}_{q^m}$, em que $\alpha^s \in \mathbb{F}_{q^m}$ e $0 \leq s < q^m$, então o conjunto ciclotômico de α^s é dado por:

$$\left\{\alpha^s, \alpha^{sq}, \alpha^{sq^2}, \ldots, \alpha^{sq^{m-1}}\right\} \mod n, \tag{2.6}$$

cujos elementos são denominados conjugados de α^s. O polinômio irredutível de menor grau sobre $\mathbb{F}_q$ com $f(\alpha^s) = 0$ é chamado de polinômio mínimo de α^s cujo grau é um divisor de m e suas raízes são α^s e seus conjugados.

***Exemplo* 2.1** *Considerando o corpo finito* $\mathbb{F}_4 = \{0, \beta, \beta^2, 1\}$ *e o polinômio primitivo* $f(x) = x^2 + x + \beta$*, os elementos do corpo de extensão* $\mathbb{F}_{16}$ *são:*

$$\begin{array}{rclrcl}
\alpha &=& \alpha & \alpha^9 &=& \beta\alpha + \beta \\
\alpha^2 &=& \alpha + \beta & \alpha^{10} &=& \beta^2 \\
\alpha^3 &=& \beta^2\alpha + \beta & \alpha^{11} &=& \beta^2\alpha \\
\alpha^4 &=& \alpha + 1 & \alpha^{12} &=& \beta^2\alpha + 1 \\
\alpha^5 &=& \beta & \alpha^{13} &=& \beta\alpha + 1 \\
\alpha^6 &=& \beta\alpha & \alpha^{14} &=& \beta^2\alpha + \beta^2 \\
\alpha^7 &=& \beta\alpha + \beta^2 & \alpha^{15} &=& 1 \\
\alpha^8 &=& \alpha + \beta^2 & &&
\end{array}$$

As classes ciclotômicas dos elementos devem ser calculadas usando a Equação (2.6). No caso do elemento α*, sua classe ciclotômica é* $\{\alpha, \alpha^4\}$ *e, portanto, o seu polinômio mínimo é* $f(x) = (x + \alpha)(x + \alpha^4) = x^2 + x + \beta$*.*

Um código linear $\mathcal{C}$ é um subespaço de $\mathbb{F}_q^n$ cujos parâmetros são $[n, k, d]_q$ em que n é o comprimento, k é a dimensão e d é a distância mínima de Hamming. Esses códigos sempre podem decodificar um vetor $\mathbf{u} \in \mathbb{F}_q^n$ de forma única se o número de erros é no máximo $t = \lfloor \frac{d-1}{2} \rfloor$. Os códigos BCH formam uma classe de códigos lineares.

Um código BCH, denotado por $\mathcal{C}_{BCH}$, é um código com parâmetros $[n, k, d]$ sobre o corpo finito $\mathbb{F}_q$, em que n é um inteiro positivo, n divide $q^m - 1$, no qual, as palavras $\mathbf{c}$ do código são os vetores:

$$\mathbf{c} = [c_0 \quad c_1 \quad \ldots \quad c_{n-1}] \in \mathbb{F}_q^n, \tag{2.7}$$

ou, em sua forma polinomial:

$$c(x) = c_0 + c_1 x + \cdots + c_{n-1} x^{n-1} \in \mathbb{F}_q[x], \tag{2.8}$$

que satisfazem,

$$c(\alpha^j) = 0 \quad \text{para} \quad j = b, b+\ell \ldots, b+\ell(d-2), \tag{2.9}$$

em que b, d e ℓ são inteiros positivos tal que $0 \leq b < n$, $1 \leq d \leq n$ e o máximo divisor comum entre ℓ e n é 1, ou seja, $\text{mdc}(\ell, n) = 1$. Além disso, α é um elemento de ordem n em $\mathbb{F}_{q^m}$. O polinômio gerador do código é dado por:

$$g(x) = \text{mmc}(f_b(x), f_{b+\ell}(x), \ldots, f_{b+\ell(d-2)}(x)), \tag{2.10}$$

em que, mmc é o mínimo múltiplo comum, $\text{grau}(g(x)) = n - k$ e $f_i(x)$ é o polinômio mínimo de α^i.

***Exemplo* 2.2** *Considerando $n = 15$ e $d = 3$, um código $\mathcal{C}_{BCH}$ sobre $\mathbb{F}_4 = \{0, \beta, \beta^2, 1\}$ pode ser definido por mais de um polinômio gerador. Por exemplo:*

(a) Se $b = 0$ e $\ell = 1$, então $g(x) = (x+1)(x+\alpha)(x+\alpha^4) = x^3 + \beta^2 x + \beta$;

(b) Se $b = 1$ e $\ell = 1$, então $g(x) = (x+\alpha)(x+\alpha^2)(x+\alpha^4)(x+\alpha^8) = x^4 + x + 1$;

(c) Se $b = 1$ e $\ell = 2$, então $g(x) = (x+\alpha)(x+\alpha^3)(x+\alpha^4)(x+\alpha^{12}) = x^4 + \beta x^3 + \beta$.

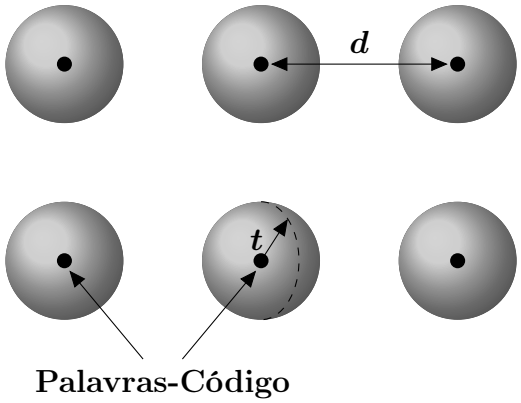
Decodificação

No contexto dos códigos corretores de erros, a decodificação é o processo de detecção e correção de erro. Portanto, decodificar uma palavra recebida significa traduzi-la em uma palavra-código. A hiperesfera de decodificação ao redor de cada palavra-código tem raio $t = \lfloor \frac{d-1}{2} \rfloor$. Nessa hiperesfera existem todos os vetores que estão a uma distância t da palavra-código para a qual esses vetores são decodificados.

Na Figura 2.1, são ilustradas as três regiões do espaço vetorial nas quais um vetor pode ocupar. A primeira região é quando o vetor é exatamente uma palavra-código; a segunda é a região sombreada, ou seja, quando o vetor está na região de decodificação (ou seja, na hiperesfera de raio t); e a terceira região é a branca, a qual consiste dos vetores que possuem distância maior que t de todas as palavras-código. Embora existam muitas técnicas para decodificação dos códigos BCH, uma vez que em nossas aplicações estaremos interessados em $\mathcal{C}_{BCH}$ com $d = 3$, por uma questão de simplicidade, será utilizado o decodificador Peterson–Gorenstein–Zierler (PGZ) [69].

Considerando $\mathcal{C}_{BCH}$ um código com parâmetros $[n, k, 3]_4$ sobre $\mathbb{F}_4$, o decodificador PGZ para esses códigos poderá, portanto, corrigir até 1 erro. Isso significa que o decodificador PGZ deve traduzir um vetor recebido, $\mathbf{r}$, em uma palavra-código válida, $\mathbf{c}$, sempre que a distância de Hamming entre $\mathbf{r}$ e $\mathbf{c}$ é 1.

Figura 2.1 – Regiões de decodificação.



Fonte: Elaborada pela autora.

Supondo que uma palavra-código $\mathbf{c} \in \mathcal{C}_{BCH}$ é transmitida por um canal ruidoso. A palavra recebida, $\mathbf{r}$, é a soma de $\mathbf{c}$ com um erro $\mathbf{e}$. Ao assumir que apenas um erro ocorreu, têm-se que, polinomialmente, $e(x) = \epsilon x^i$, em que i é a posição desconhecida do erro e ϵ é magnitude. Uma vez que as raízes do polinômio gerador são também raízes da palavra-código, ao avaliar $\mathbf{r}$ nessas raízes obtém-se as síndromes que devem isolar o erro. As síndromes são calculadas da seguinte maneira:

$$s_j = r(\alpha^j) = c(\alpha^j) + e(\alpha^j), \tag{2.11}$$

em que, $j = b,\ b + \ell$. Como α^j são raízes de $g(x)$, então $c(\alpha^j) = 0$. O problema de decodificação é então reduzido a determinar a solução do seguinte conjunto de equações não lineares:

$$\begin{cases} s_b &= \epsilon\alpha^{bi} \\ s_{b+\ell} &= \epsilon\alpha^{(b+\ell)i} = s_b\alpha^i \end{cases}. \tag{2.12}$$

Se todas as síndromes são zero, então não houve erro e a decodificação está concluída. Caso contrário, a posição do erro é determinada pela solução segunda equação do sistema não linear da Equação (2.12), e a magnitude do erro é dada pela solução da primeira equação. O decodificador, portanto, retorna uma palavra-código válida se $0 \leq i < n$ e $\epsilon \in \mathbb{F}_q$. Se apenas uma síndrome é zero, então o número de erros deverá ter excedido a capacidade de correção de erros do código $\mathcal{C}_{BCH}$.

***Exemplo* 2.3** *Considerando que uma palavra-código* $\mathbf{c} \in \mathcal{C}_{BCH}$ *sobre* $\mathbb{F}_4 = \{0, \beta, \beta^2, 1\}$ *com comprimento* $n = 15$*, distância* $d = 3$ *e polinômio gerador* $g(x) = x^3 + \beta^2 x + \beta$ *foi transmitida por um canal ruidoso. Supondo que a palavra recebida foi:*

$$\mathbf{r} = [\beta \ \ 1 \ \ \beta^2 \ \ 0 \ \ 1 \ \ \beta \ \ \beta \ \ 1 \ \ 1 \ \ \beta \ \ 0 \ \ 0 \ \ \beta^2 \ \ \beta \ \ \beta^2],$$

ou seja,

$$r(x) = \beta^2 x^{14} + \beta x^{13} + \beta^2 x^{12} + \beta x^9 + x^8 + x^7 + \beta x^6 + \beta x^5 + x^4 + \beta^2 x^2 + x + \beta.$$

Usando a aritmética de $\mathbb{F}_{16}$ *para calcular as síndromes, têm-se que:* $s_0 = 1$ *e* $s_1 = \alpha^{12}$. *Resolvendo o sistema da Equação (2.12), sabe-se que o erro ocorreu na décima segunda componente pois* $\alpha^i = \frac{s_1}{s_0} = \alpha^{12}$, *então* $i = 12$; *e a magnitude é* $\epsilon = 1$, *assim,* $e(x) = x^{12}$. *Por fim, sabe-se que a palavra-código transmitida foi:*

$$\mathbf{c} = [\beta \;\; 1 \;\; \beta^2 \;\; 0 \;\; 1 \;\; \beta \;\; \beta \;\; 1 \;\; 1 \;\; \beta \;\; 0 \;\; 0 \;\; \beta \;\; \beta \;\; \beta^2].$$

2.2 Análise Espectral

Em muitas aplicações do mundo real, os sinais representados no domínio do tempo são incapazes de inferir as informações e padrões ocultos no sinal. Portanto, é necessário representar o sinal em alguns domínios alternativos, nos quais as características intrínsecas ao sinal podem ser expressas de outra maneira, melhorando, assim, sua interpretabilidade. A transformada de Fourier, por exemplo, é a operação que representa um sinal originalmente do domínio do tempo no domínio da frequência. Nesse contexto, a transformada de Fourier tem sido uma das técnicas de análise espectral mais comuns.

2.2.1 Transformada de Fourier para Sinais no Tempo Discreto

Considera-se um sinal aperiódico discreto no tempo $x[n]$ de duração finita N. A representação no domínio da frequência a partir da transformada de Fourier para sinais no tempo discreto (DFT, do inglês: *Discrete Fourier Transform*) é uma função periódica que expressa um sinal aperiódico $x[n]$ de duração finita N como uma combinação linear de exponenciais complexas $e^{j\Omega n}$. A frequência $\Omega = k\frac{2\pi}{N}$ assume valores em um intervalo contínuo de tamanho 2π. A equação de síntese, ou transformada direta, é dada por:

$$\mathcal{F}\left[x[n]\right] = X[k] = \sum_{n=1}^{N} x[n] e^{-jk\frac{2\pi}{N}n}, \tag{2.13}$$

em que, $\mathcal{F}$ é o operador transformada de Fourier no tempo discreto [70, 71]. O cálculo direto dos coeficientes da DFT requer $O(N^2)$ operações. Porém, existe um algoritmo eficiente, conhecido como transformada rápida de Fourier (FFT, do inglês: *Fast Fourier Transform*), que reduz a complexidade computacional para $O(N \log_2 N)$ operações.

A partir da representação no domínio da frequência é possível fazer a análise do espectro de energia do sinal, e, assim, analisar as periodicidades que compõem o sinal. O espectro de energia descreve como a energia do sinal está distribuída ao longo do intervalo de frequências Ω e é definido em função da DFT como:

$$S[k] = |X[k]|^2. \tag{2.14}$$

2.2.2 Transformada de Fourier de Tempo Reduzido

Ao contrário da DFT, que é uma representação em frequência de um sinal discreto, a transformada de Fourier de tempo reduzido (STFT, do inglês: *Short-time Fourier Transform*) é uma representação nos domínios do tempo e da frequência simultaneamente. A análise de tempo-frequência é de grande interesse quando os modelos de sinal não estão disponíveis. Em tais casos, ter apenas uma das representações do sinal, seja no domínio do tempo ou no domínio da frequência, não é suficiente para extrair informações para classificação de características [72].

A STFT identifica as componentes de frequência de um sinal ao longo do tempo. Uma janela de comprimento W é deslizada por toda a extensão do sinal recalculando a transformada de Fourier para cada janela. Ao considerar um sinal discreto no tempo $x[n]$, a sua STFT é dada por:

$$STFT[\tau, k] = \sum_{n=1}^{N} x[n]\mathrm{w}[n-\tau]e^{-jk\frac{2\pi}{N}n}, \tag{2.15}$$

em que, $\mathrm{w}[n-\tau]$ é a função de janela de comprimento W que está centrada em diferentes instantes de tempo, τ, no plano tempo-frequência (essas janelas podem, ou não, se sobrepor) [73]. Alguns exemplos de funções de janela são apresentados na Figura 2.2.

A STFT também pode ser interpretada como um filtro digital sendo recorrentemente aplicado ao sinal $x[n]$. Nessa transformada, é fixada uma resolução para o tempo e para frequência, em que ambas dependem do comprimento da função de janela. Assim, uma janela curta é necessária para uma boa resolução de tempo e uma janela mais ampla oferece uma boa resolução de frequência.

2.2.3 Filtragem Digital

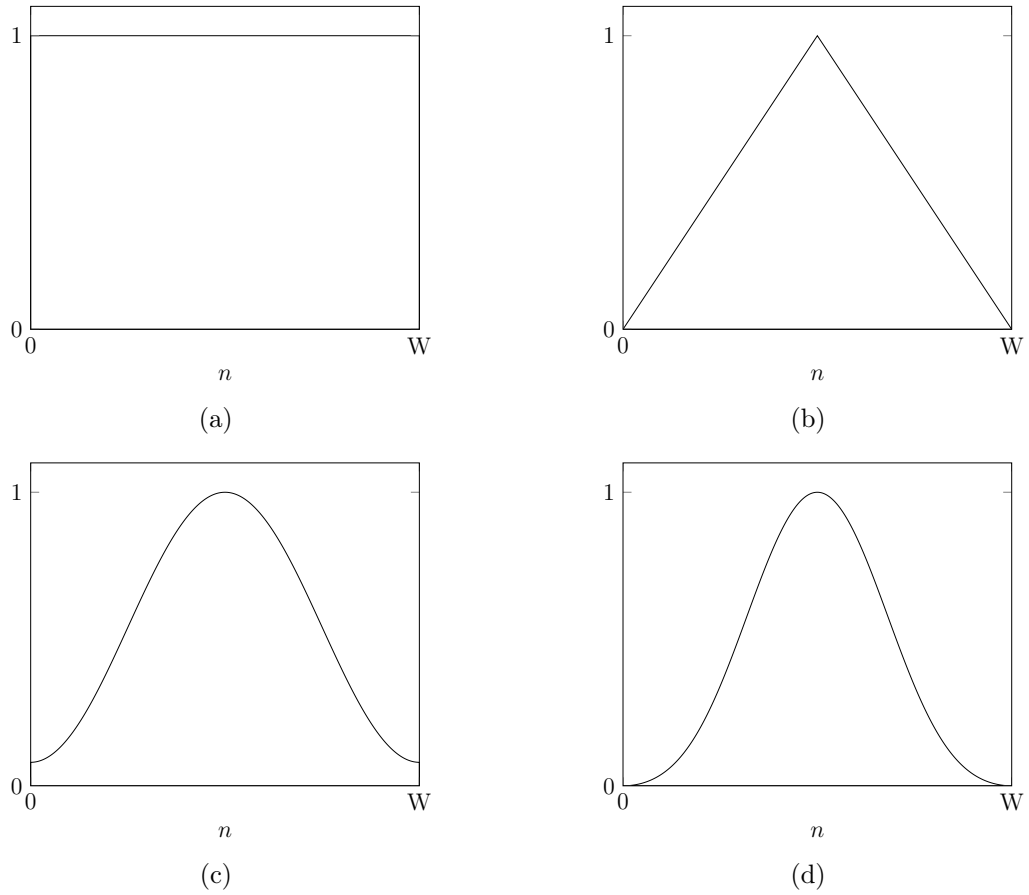
Um filtro é um sistema linear invariante no tempo usado para realizar uma filtragem seletiva de frequência. Um filtro é aplicado no processamento digital de sinais de várias maneiras. Por exemplo, para remoção de ruído indesejável e para análise espectral de sinais. Os filtros são classificados de acordo com suas características no domínio da frequência como passa-baixa, passa-alta, passa-faixa e rejeita-faixa.

Em aplicações em que uma frequência específica deve ser eliminada, são aplicados, por exemplo, os filtros *notch* [74]. Os filtros *notch* são rejeita-faixa e os filtros *notch* complementares são passa-faixa com largura de banda estreita. O projeto de dos filtros *notch* considera, inicialmente, um filtro passa-alta de segunda ordem definido no domínio $z = e^{j\Omega}$ como:

$$A[z] = \frac{r^2 - 2r\cos w_0 z^{-1} + z^{-2}}{1 - 2r\cos w_0 z^{-1} + r^2 z^{-2}}, \tag{2.16}$$

em que, r é o parâmetro de ajuste de banda do filtro [55, 9]. Esses filtros são sensíveis a pequenas mudanças de r. Definindo dois filtros $G[z]$ e $H[z]$ obtidos a partir de $A[z]$ como:

Figura 2.2 – Funções de janela: (a) Retangular; (b) Bartlett; (c) Hamming; e (d) Blackman.



Fonte: Elaborada pela autora.

$$\begin{bmatrix} G[z] \\ H[z] \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ A[z] \end{bmatrix}. \tag{2.17}$$

Têm-se que,

$$G[z] = \frac{1+r^2}{2} \left[\frac{1 - 2\cos w_0 z^{-1} + z^{-2}}{1 - 2r\cos w_0 z^{-1} + r^2 z^{-2}} \right], \tag{2.18}$$

será um filtro *notch* quando o r se aproximar de 1. Além disso, $G[z]$ e $H[z]$ são complementares, assim,

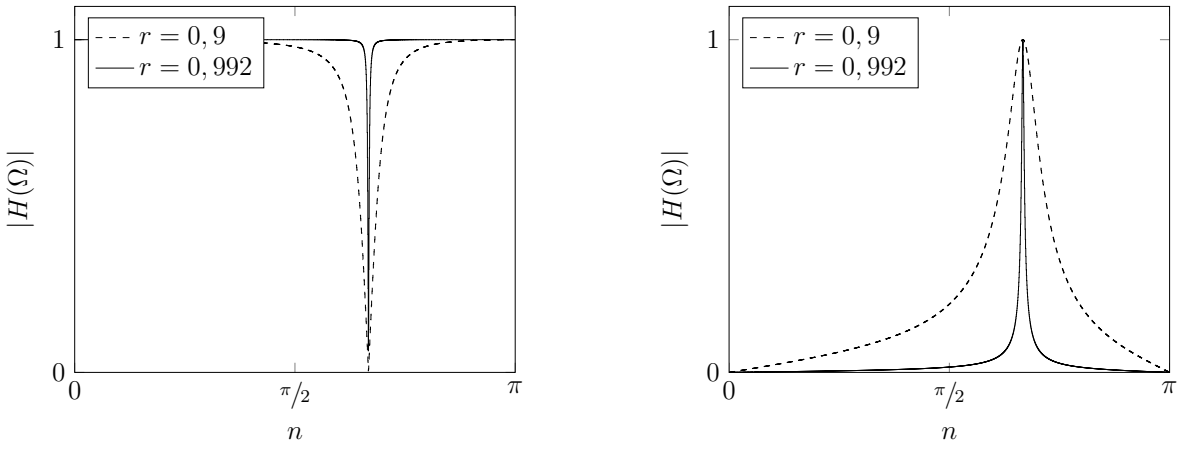
$$H[z] = \frac{1}{2} \left[\frac{(1-r^2)(1-z^{-2})}{1 - 2r\cos w_0 z^{-1} + r^2 z^{-2}} \right]. \tag{2.19}$$

será um filtro *notch* complementar. A magnitude da resposta em frequência dos filtros $G[z]$ e $H[z]$ para $w_0 = {}^{2\pi}/_{3}$ é apresentada na Figura 2.3 para dois valores distintos de r.

2.3 Complexidade de Sequências Finitas

A aleatoriedade de uma sequência finita está relacionada com a dificuldade em prever o próximo dígito da sequência a partir dos símbolos anteriores. Intuitivamente, uma

Figura 2.3 – Magnitude da resposta em frequência para $r = 0,9$ e $r = 0,992$ dos filtros: (a) *notch*; e (b) *notch* complementar.



Fonte: Elaborada pela autora.

sequência aleatória deve carecer de qualquer regularidade, de modo que cada novo símbolo da sequência seja difícil de prever. A dificuldade de previsão significa que a probabilidade de predizer corretamente o próximo símbolo é menor ou igual a de predizer erroneamente. Por exemplo, considerando as seguintes sequências cujos comprimentos são $n = 40$ e o alfabeto seja binário:

$$\begin{aligned} s_1 &= 0000000000000000000000000000000000000000, \\ s_2 &= 0110011001100110011001100110011001100110, \\ s_3 &= 1001000000000011000000000101001100000001, \\ s_4 &= 0101110001101001101010110101101110010110. \end{aligned}$$

Embora cada uma das sequências tenha probabilidade 2^{-40} de ser escolhida aleatoriamente em um espaço de dimensão $n = 40$, as quatro sequências são diferentes entre si. A sequência s_1, por exemplo, é composta apenas por símbolos zero e s_2 é a repetição da subsequência 0110. Nesse caso, ambas as sequências s_1 e s_2 tem padrões regulares e, portanto, são previsíveis. Porém, em s_3 e s_4 não existem padrões regulares óbvios, o que significa que há alguma dificuldade em prever o próximo símbolo.

Uma abordagem comumente usada para medir essa aleatoriedade é considerar uma medida de complexidade. Porém, não existe uma medida única para isso. Ray Solomonoff [75] e Andrey Kolmogorov [76] usaram, independentemente, a "ausência de padrões" de uma sequência finita, determinada pelo comprimento do menor programa para máquina de Turing que a gerou, para medir a complexidade dessa sequência. Assim, uma vez que essa complexidade independe da distribuição de probabilidade dos símbolos da sequência, então a descrição mais curta do computador atua como um código universal que é uniformemente bom para todas as sequências [77].

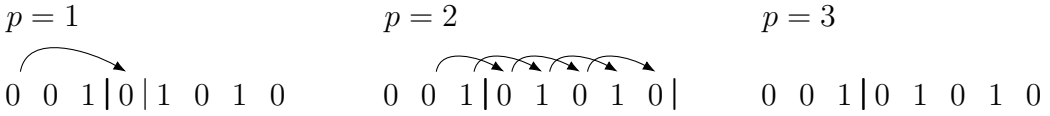
Para estimar a complexidade de Kolmogorov utiliza-se algoritmos que refletem a capacidade de representar uma sequência de forma compacta com base em suas características estruturais. Uma abordagem amplamente discutida para esse contexto foi proposta por Abraham Lempel e Jacob Ziv [78]. Eles caracterizaram a complexidade de uma sequência finita pelo número de etapas em um processo de cópia recursiva de partes de uma sequência.

2.3.1 Complexidade de Lempel-Ziv

A complexidade proposta por Abraham Lempel e Jacob Ziv [78], comumente chamada de complexidade de Lempel-Ziv, tem sido aplicada, principalmente, para compressão de dados [79, 80, 81]. Em tal método, a complexidade de uma sequência finita é calculada do ponto de vista de uma máquina de aprendizagem autodeterminada. Assim, uma sequência $s = s_1 s_2 \dots s_n$ é verificada da esquerda para a direita e uma nova palavra é adicionada a seu dicionário a cada vez que é descoberta uma subsequência de dígitos consecutivos não encontrados anteriormente. O tamanho do dicionário e a taxa com que novas palavras são encontradas ao longo de s são fundamentais para calcular a complexidade de Lempel-Ziv.

O mecanismo pelo qual a complexidade de Lempel-Ziv de s é avaliada utiliza três conceitos: reprodutibilidade, produtibilidade e história exaustiva. Uma sequência s de comprimento n é dita reproduzível por seu prefixo s^j, e denotado por $s^j \to s$, quando s_{j+1}^n é uma subsequência de s^{n-1}, sendo assim, $s_{j+1}^n = s_p^{n-j+p-1}$ para algum $p \leq j$.

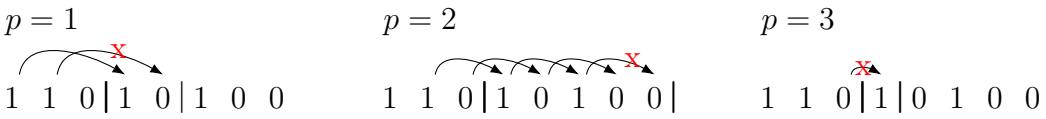
***Exemplo* 2.4** *Considerando a sequência* $s = 00101010$ *deseja-se verificar se o prefixo* 001 *reproduz* s*. A reprodutibilidade deve ser verificada para cada ponteiro* $p \leq j$*, e, portanto, para esse exemplo,* $p \leq 3$*. Assim, iterativamente, essa análise é feita conforme a seguir:*



Logo, a sequência não é reproduzida completamente para os ponteiros $p = 1$ *e* $p = 3$*, mas, para* $p = 2$*,* $001 \to 00101010$*.*

Por outro lado, uma sequência s é produzida por seu prefixo s^j, e denotado por $s^j \Rightarrow s$, quando $s^j \to s^{n-1}$, ou seja, quando s_{j+1}^{n-1} é uma subsequência de s^{n-2}, sendo assim, $s_{j+1}^{n-1} = s_p^{n-j+p-2}$ para algum $p \leq j$.

***Exemplo* 2.5** *Considerando a sequência* $s = 11010100$ *deseja-se verificar qual o ponteiro que permite produzir a sequência mais longa para o prefixo* 110*. A produtibilidade deve ser verificada para cada ponteiro* $p \leq j$*, e, portanto, para esse exemplo,* $p \leq 3$*. Assim, iterativamente, essa análise é feita conforme a seguir:*
Logo, para $p = 1$*,* $110 \Rightarrow 11010$*, para* $p = 2$*,* $110 \Rightarrow 11010100$ *e para* $p = 3$*,* $110 \Rightarrow 1101$*.*



A diferença entre produtibilidade e reprodutibilidade é que a primeira permite um símbolo extra diferente no final do processo de extensão, o que não é permitido no último. Portanto, uma sequência que é reproduzível por um prefixo, é, também, sempre produzida por ele, contudo, o inverso nem sempre é verdadeiro.

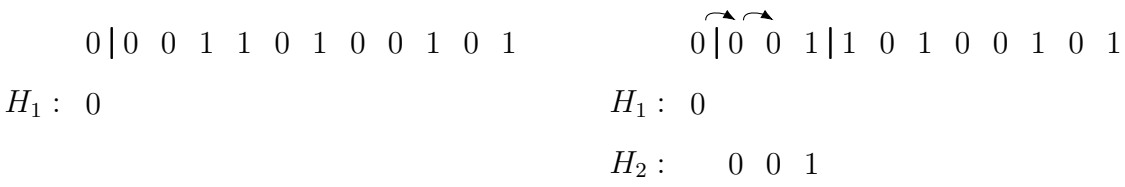
***Exemplo* 2.6** *Considerando a sequência* $s = 0100$*, com relação ao prefixo* 01*, têm-se que:* $01 \Rightarrow 0100$ *para* $p = 1$*, mas* $01 \not\Rightarrow 0100$*, em que* $\not\Rightarrow$ *é a negação de* $\rightarrow$*.*

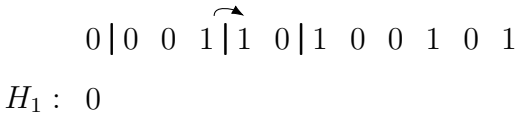
De fato, toda sequência não nula s pode ser produzida por algum prefixo adequado. Assim, qualquer sequência finita não nula e de comprimento n pode ser interpretada como representando o produto final de um processo recursivo de construção de dicionário auto-delimitado. Na primeira etapa, a subsequência vazia $\Lambda \Rightarrow s_1$, então, por um processo recursivo, na etapa i já foi produzido s^{h_i} e segue que $s^{h_i} \Rightarrow s^{h_{i+1}}$, e assim por diante. Após no máximo n etapas, toda sequência s foi produzida. Considerando $1 \leq m \leq n$ o número de passos necessários para este processo, a história de s corresponde à seguinte decomposição:

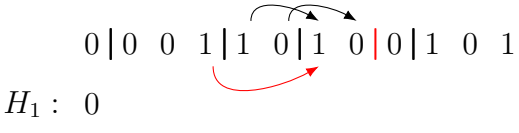
$$H(s) = s_1^{h_1} s_{h_1+1}^{h_2} s_{h_2+1}^{h_3} \cdots s_{h_{m-1}+1}^{h_m}, \tag{2.20}$$

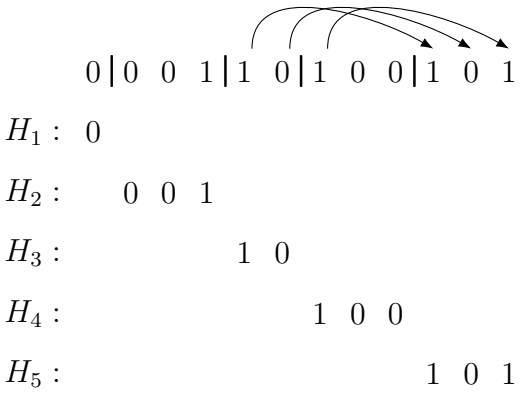
em que $h_1 = 1$, $h_m = n$ e $H_i(s) = s_{h_{i-1}+1}^{h_i}$ para $i = 1, 2, \ldots, m$ são as componentes de $H(s)$. Uma história de s é considerada exaustiva se $s^{h_{i-1}} \Rightarrow s^{h_i}$ mas $s^{h_{i-1}} \not\Rightarrow s^{h_i}$. Qualquer sequência finita não nula tem apenas uma história exaustiva, e essa história tem o menor número de componentes de todas as histórias possíveis de s [78]. O número de componentes da história exaustiva de s é chamada de complexidade de Lempel-Ziv de s e denotada por $\mathcal{C}_{LZ}(s)$.

***Exemplo* 2.7** *As componentes exaustivas da sequência* $s = 000110100101$ *são calculadas, iterativamente, da seguinte maneira:*









Assim, a história exaustiva é $H(s) = 0|001|10|100|101$*, em que as componentes sucessivas são separadas por barras e onde a ausência de uma barra no final da sequência indica que a última componente não é exaustiva. Mesmo assim,* $\mathcal{C}_{LZ}(s) = 5$*.*

Abraham Lempel e Jacob Ziv [78] também demonstraram que para uma sequência não nula s de comprimento n definida sobre um alfabeto $\mathcal{A}$ de cardinalidade α, essa complexidade é limitada superiormente por:

$$\mathcal{C}_{LZ}(s) \leq \frac{n}{(1-\varepsilon_n)\log_\alpha n}, \tag{2.21}$$

em que

$$\varepsilon_n = 2\frac{1+\log_\alpha \log_\alpha \alpha n}{\log_\alpha n}, \tag{2.22}$$

que é uma função de n que decai lentamente Assintoticamente, a Equação (2.21) tende a

$$\mathcal{C}_{LZ}(s) \leq \frac{n}{\log_\alpha n}. \tag{2.23}$$

Capítulo 3

Processamento de Sinal Genômico

A análise, processamento e uso de sinais genômicos para obtenção de conhecimento biológico constituem o domínio deste capítulo. O objetivo é integrar a teoria e os métodos de processamento de sinais para extração de informações genômicas. Portanto, a representação numérica e gráfica de sequências de DNA serão discutidas neste capítulo. Além disso, serão abordadas algumas técnicas de extração de informações genômicas úteis para análise espectral.

3.1 Representação Numérica

A conversão de sequências genômicas da forma simbólica, tal como fornecida nos bancos de dados genômicos públicos, em sinais digitais permite o uso de técnicas de processamento de sinal para processamento e análise de dados genômicos. Para isto, as bases de uma sequência de DNA, cujo alfabeto é $\mathcal{N} = \{\texttt{A},\ \texttt{C},\ \texttt{G},\ \texttt{T}\}$, devem ser mapeadas adequadamente para valores numéricos correspondentes. Assim, considerando que s seja uma sequência de DNA com N bases, ou seja,

$$s = s_1 s_2 \dots s_N, \tag{3.1}$$

em que $s_i \in \mathcal{N},\ \forall i \in [1, N]$, um mapeamento $\mathcal{M}$ é definido como a regra de associação entre as quatro bases de DNA e quatro números complexos, isto é,

$$\mathcal{M}:\ \texttt{A} \mapsto a,\ \texttt{C} \mapsto c,\ \texttt{G} \mapsto g,\ \texttt{T} \mapsto t, \tag{3.2}$$

tal que, sua imagem é dada por,

$$a,\ c,\ g,\ t \in \mathbb{C}. \tag{3.3}$$

Assim, supondo que as quatro primeiras bases de uma dada sequência de DNA sejam `ACGT`, podemos, portanto, associá-la ao seguinte sinal discreto no tempo usando $\mathcal{M}$:

$$x[n] = a\delta[n] + c\delta[n-1] + g\delta[n-2] + t\delta[n-3] + \cdots, \tag{3.4}$$

em que $\delta[n]$ é o sinal impulso unitário discreto, definido como

$$\delta[n] = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} . \tag{3.5}$$

Uma maneira alternativa para escrever a Equação (3.4) é,

$$x[n] = ax_{\mathtt{A}}[n] + cx_{\mathtt{C}}[n] + gx_{\mathtt{G}}[n] + tx_{\mathtt{T}}[n], \tag{3.6}$$

em que $x_\alpha[n]$ é a função indicador binário. A função indicador binário assume 1 quando o n-ésimo símbolo de s é uma base $\alpha \in \{\mathtt{A},\ \mathtt{C},\ \mathtt{G},\ \mathtt{T}\}$ e 0 caso contrário.

Devido ao fato de que sinais de tempo discreto têm as mesmas propriedades básicas de vetores, um dado sinal $x[n]$ definido em algum intervalo $[0, N[$ pode ser representado na forma vetorial como $\mathbf{x} = [x_0 \ \ x_1 \ \ \ldots \ \ x_{N-1}]$. Assim, a Equação (3.6) ainda pode ser reescrita na forma vetorial da seguinte maneira:

$$\begin{aligned} \mathbf{x} &= a\mathbf{x}_{\mathtt{A}} + c\mathbf{x}_{\mathtt{C}} + g\mathbf{x}_{\mathtt{G}} + t\mathbf{x}_{\mathtt{T}} \\ &= \mathbf{w} \begin{bmatrix} \mathbf{x}_{\mathtt{A}} \\ \mathbf{x}_{\mathtt{C}} \\ \mathbf{x}_{\mathtt{G}} \\ \mathbf{x}_{\mathtt{T}} \end{bmatrix}, \end{aligned} \tag{3.7}$$

em que $\mathbf{w} = [a \ \ c \ \ g \ \ t]$ são os elementos de $\mathcal{M}$ e é denominada vetor de peso. Cada vetor correspondente a uma função indicador binário é N-dimensional. Por fim, $\mathbf{x}$ é também N-dimensional.

***Exemplo* 3.1** *As funções indicador binário para a sequência* $s = \mathtt{CTGATCCTTCAAGCG}$ *em sua representação vetorial são dadas por:*

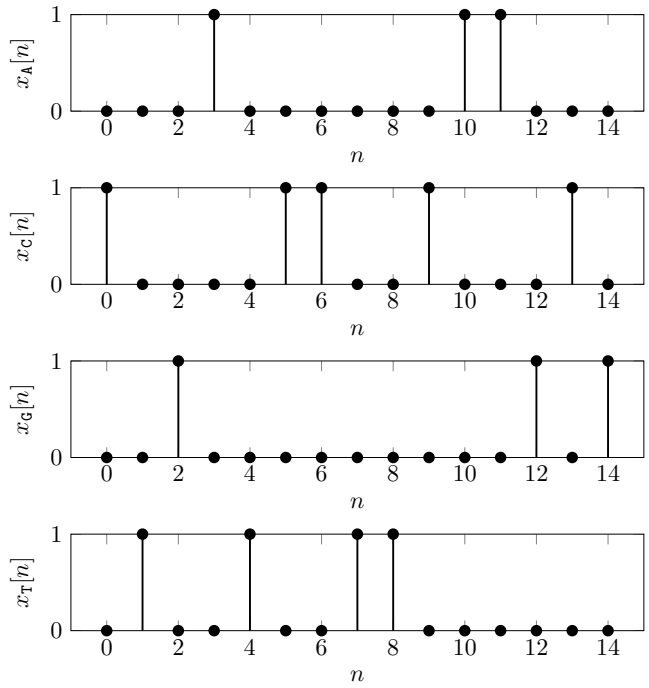
$$\begin{aligned} \mathbf{x}_{\mathtt{A}} &= [0 \ \ 0 \ \ 0 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 1 \ \ 1 \ \ 0 \ \ 0 \ \ 0], \\ \mathbf{x}_{\mathtt{C}} &= [1 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 1 \ \ 1 \ \ 0 \ \ 0 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 1 \ \ 0], \\ \mathbf{x}_{\mathtt{G}} &= [0 \ \ 0 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 1 \ \ 0 \ \ 1], \\ \mathbf{x}_{\mathtt{T}} &= [0 \ \ 1 \ \ 0 \ \ 0 \ \ 1 \ \ 0 \ \ 0 \ \ 1 \ \ 1 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0]. \end{aligned}$$

Além disso, sua representação gráfica é mostrada na Figura 3.1. É possível observar que o conjunto de sinais indicadores binários somam 1 *para todo* n.

Esquemas de Mapeamento

Existem diversos esquemas de mapeamento encontrados na literatura. Eles podem ser amplamente classificados em quatro categorias: unidimensional, multidimensional, cumulativo e adaptativo. Para cada uma dessas categorias, existem mais de um esquema de mapeamento. Os mais comuns estão resumidos na Tabela 3.1 e serão descritos a seguir.

Figura 3.1 – Funções indicador binário para $s = \texttt{CTGATCCTTCAAGCG}$.



Fonte: Elaborada pela autora.

O mapeamento definido na Equação (3.2) é característico do mapeamento unidimensional, cuja única restrição é que os números a, c, g e t sejam distintos. No caso do mapeamento multidimensional, existem dois ou mais sinais $x[n]$ em que, para cada um deles, são definidos valores de a, c, g e t. Já no mapeamento cumulativo (que pode ser unidimensional ou multidimensional), a, c, g e t são atualizados em função de n. Por fim, no mapeamento adaptativo, a, c, g e t são unicamente determinados para sequência baseado em suas características intrínsecas.

Os mapeamentos unidimensionais são caracterizados pelo mapeamento bijetivo de cada uma das quatro bases para um único valor numérico. Destacam-se:

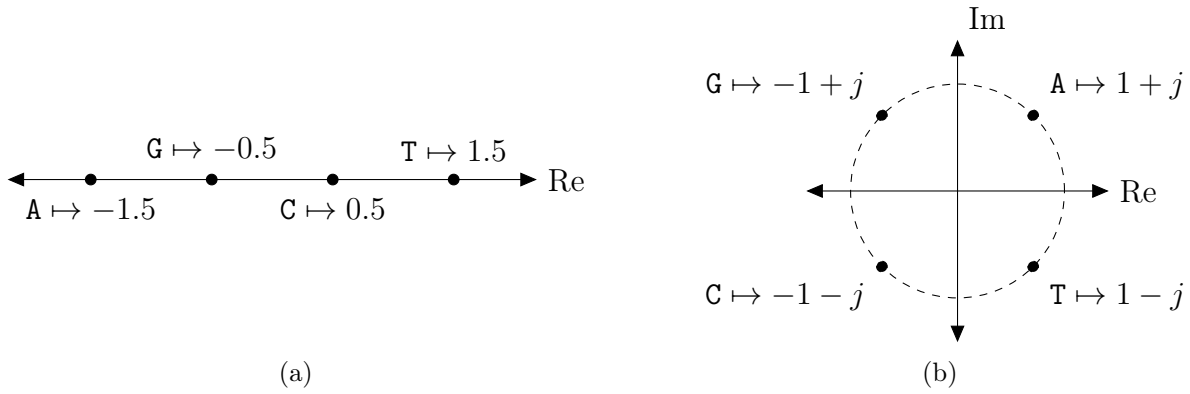
- Inteiro: Cristea sugeriu que, embora existam 24 escolhas de representação considerando os números inteiros $\{0, 1, 2, 3\}$, o mapeamento ideal deve ter o melhor sinal auto-correlacionado e é dado por: $\texttt{A} \mapsto 2$, $\texttt{C} \mapsto 1$, $\texttt{G} \mapsto 3$ e $\texttt{T} \mapsto 0$ [82];

- Real: Chakravarthy *et al.* propuseram a seguinte regra de mapeamento para número real baseada na propriedade complementar do DNA: $\texttt{A} \mapsto -1.5$, $\texttt{C} \mapsto 0.5$, $\texttt{G} \mapsto -0.5$ e $\texttt{T} \mapsto 1.5$ [83];

- Complexo: Anastassiou também analisou a propriedade complementar do DNA para propor o seguinte mapeamento: $\texttt{A} \mapsto 1 + j$, $\texttt{C} \mapsto -1 - j$, $\texttt{G} \mapsto -1 + j$ e $\texttt{T} \mapsto 1 - j$ [5];

- Potencial de interação elétron-íon (EIIP, do inglês: *electron-ion interaction pseudopotential*): Lalović *et al.* sugeriu empregar valores numéricos que representam

Figura 3.2 – Diagrama de constelação para as representações (a) real e (b) complexo.



(a) (b)

Fonte: Elaborada pela autora

a distribuição das energias do elétron livre ao longo da sequência de DNA, assim: $\texttt{A} \mapsto 0.1260$, $\texttt{C} \mapsto 0.1340$, $\texttt{G} \mapsto 0.0806$ e $\texttt{T} \mapsto 0.1335$ [54].

As representações inteiras, reais e complexas também podem ser interpretadas como os diagramas de constelação amplamente usados em comunicações digitais. Na Figura 3.2, os diagramas de constelação para as representações reais e complexas são apresentados. A constelação do mapeamento real é semelhante ao esquema de modulação por amplitude de pulso (PAM, do inglês: *Pulse Amplitude Modulation*), e a do mapeamento complexo é semelhante ao esquema de modulação QPSK [83].

No mapeamento multidimensional destacam-se:

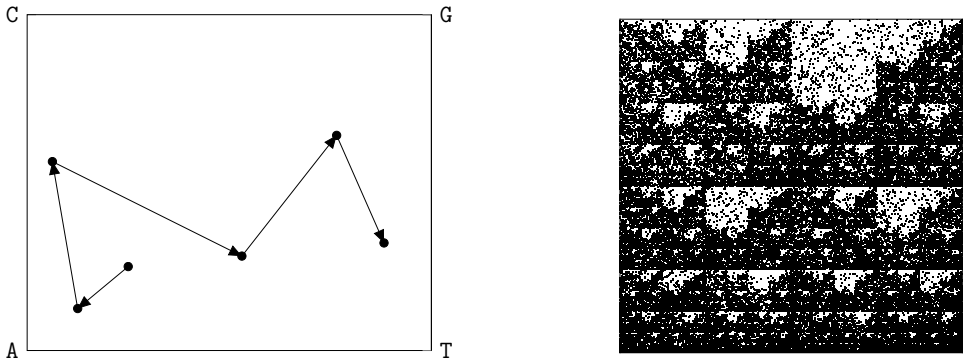
- Indicador binário: Voss propôs o mapeamento em quatro sequências binárias para indicar a presença com bit 1 ou ausência com bit 0 do respectivo nucleotídeo [53];
- Tetraedro: Silverman *et al.* propuseram que cada uma das quatro bases seja atribuída a um vértice do tetraedro regular no espaço, reduzindo, assim, o número de sequências indicadoras de quatro para três de maneira simétrica para todos os quatro componentes [84].

As representações cumulativas podem ser unidimensionais ou multidimensionais e são caracterizadas pelo emprego de um modelo de passeio aleatório no qual uma curva é construída pela contribuição agregada de valores numéricos consecutivos atribuídos a cada base. Destaca-se:

- *Chaos game representation* (CGR): Jeffrey se baseou em uma técnica de sistemas dinâmicos caóticos para propor um método que produz uma imagem, em que cada ponto está exatamente no meio com relação ao ponto anterior e o vértice correspondente à base atual [85].

A imagem gerada pelo CGR é conhecida como atrator, que revela alguma estrutura subjacente em uma sequência de DNA. Cada nucleotídeo é representado por um par

Figura 3.3 – *Chaos ame representation* para: (a) $s =$ `AACTGT`; e (b) Beta globina humana, HBB com 73308 bp.



(a) (b)

Fonte: Elaborada pela autora

ordenado que corresponde a uma posição no quadrado unitário. Para a sequência tomada como exemplo na Tabela 3.1, $s =$ `AACTGT`, o CGR é construído conforme a imagem da Figura 3.4(a). Além disso, na Figura 3.4(b) é apresentado o mapeamento CGR para o gene HBB com 73308 bp.

No contexto do mapeamento adaptativo, cada sequência de DNA deve ser mapeada para um sinal numérico conforme algum critério pré estabelecido. Essa abordagem é essencialmente importante, uma vez que um único mapeamento para qualquer sequência de DNA pode está ignorando propriedades intrínsecas de tais sequências. Destaca-se:

- Mapeamento de Entropia Mínima (MEM, do inglês: *Minimum Entropy Mapping*): Galleani *et al.* [8] propuseram um algoritmo no qual um mapeamento adaptativo e real é calculado individualmente para cada sequência considerando o critério de minimização da entropia espectral.

Cada representação numérica do DNA detém propriedades diferentes. No caso da representação por indicador binário, nenhuma relação de ordem é imposta às bases do DNA. Por essa razão, esse mapeamento tem sido amplamente usado para análise espectral de sequências de DNA [4, 5, 14]. Além disso, o indicador binário, o tetraedro e CGR são representações multidimensionais que introduzem redundâncias em cada vetor unidimensional.

As representações em números inteiros e reais podem introduzir algumas propriedades matemáticas que não existem em uma sequência genômica, portanto, esses mapeamentos devem ser usados com cuidado à depender da aplicação. O mapeamento para os números complexos tem semelhanças com o esquema QPSK de modulação digital e reflete a característica complementar do DNA. Já o EIIP reflete as propriedades físico-química e pode melhorar a capacidade de localização de genes, além de reduzir a sobrecarga computacional.

Tabela 3.1 – Representações numéricas para sequências genômicas.

		Mapeamento	$s = \texttt{AACTGT}$
i	Inteiro	$x[n] = \begin{cases} 2, & s_n = \texttt{A} \\ 1, & s_n = \texttt{C} \\ 3, & s_n = \texttt{G} \\ 0, & s_n = \texttt{T} \end{cases}$	$x[n] = [2 \;\; 2 \;\; 1 \;\; 0 \;\; 3 \;\; 0]$
ii	Real	$x[n] = \begin{cases} -1,5, & s_n = \texttt{A} \\ 0,5, & s_n = \texttt{C} \\ -0,5, & s_n = \texttt{G} \\ 1,5, & s_n = \texttt{T} \end{cases}$	$x[n] = [-1,5 \;\; -1,5 \;\; 0,5 \;\; 1,5 \;\; -0,5 \;\; 1,5]$
iii	Complexo	$x[n] = \begin{cases} 1+j, & s_n = \texttt{A} \\ -1-j, & s_n = \texttt{C} \\ -1+j, & s_n = \texttt{G} \\ 1-j, & s_n = \texttt{T} \end{cases}$	$x[n] = [1+j \;\; 1+j \;\; -1-j \;\; 1-j \;\; -1+j \;\; 1-j]$
iv	EIIP	$x[n] = \begin{cases} 0,1260, & s_n = \texttt{A} \\ 0,1340, & s_n = \texttt{C} \\ 0,0806, & s_n = \texttt{G} \\ 0,1335, & s_n = \texttt{T} \end{cases}$	$x[n] = [0,1260 \;\; 0,1260 \;\; 0,1340 \;\; 0,1335 \;\; 0,0806 \;\; 0,1335]$
v	Indicador Binário	$x_k[n] = \begin{cases} 1, & s_n = k \\ 0, & c.c. \end{cases}$ $\forall\ k \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$	$x_{\texttt{A}}[n] = [1 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0]$ $x_{\texttt{C}}[n] = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0]$ $x_{\texttt{G}}[n] = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0]$ $x_{\texttt{T}}[n] = [0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 1]$
vi	Tetraedro	$x[n] = \begin{cases} \frac{2\sqrt{2}}{3}, & s_n = \texttt{T} \\ -\frac{\sqrt{2}}{3}, & s_n = \texttt{C} \text{ ou } \texttt{G} \\ 0, & c.c. \end{cases}$ $y[n] = \begin{cases} \frac{\sqrt{6}}{3}, & s_n = \texttt{C} \\ -\frac{\sqrt{6}}{3}, & s_n = \texttt{G} \\ 0, & c.c. \end{cases}$ $z[n] = \begin{cases} 1, & s_n = \texttt{A} \\ -\frac{1}{3}, & c.c. \end{cases}$	$x[n] = [0 \;\; 0 \;\; -\frac{\sqrt{2}}{3} \;\; \frac{2\sqrt{2}}{3} \;\; -\frac{\sqrt{2}}{3} \;\; \frac{2\sqrt{2}}{3}]$ $y[n] = [0 \;\; 0 \;\; \frac{\sqrt{6}}{3} \;\; 0 \;\; -\frac{\sqrt{6}}{3} \;\; 0]$ $z[n] = [1 \;\; 1 \;\; -\frac{1}{3} \;\; -\frac{1}{3} \;\; -\frac{1}{3} \;\; -\frac{1}{3}]$
vii	CGR	$v[n] = \begin{cases} (0,0) & s_n = \texttt{A} \\ (0,1) & s_n = \texttt{C} \\ (1,1) & s_n = \texttt{G} \\ (1,0) & s_n = \texttt{T} \end{cases}$ $x[0] = y[0] = 0,5$ $x[n] = 0,5(x[n-1] + v_0[n])$ $y[n] = 0,5(y[n-1] + v_1[n])$	$x[n] = [0,25 \;\; 0,125 \;\; 0,0625 \;\; 0,5312 \;\; 0,7656 \;\; 0,8828]$ $y[n] = [0,25 \;\; 0,125 \;\; 0,5625 \;\; 0,2812 \;\; 0,6406 \;\; 0,3203]$

3.2 Análise Espectral

A DFT e o espectro de energia são comumente aplicados para análise espectral de sequências genômicas com o objetivo de distinguir sequências codificantes (CDS, do inglês: *coding sequence*) e não codificantes de proteínas [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. A busca por regularidades nas sequências de DNA mostrou a existência da periodicidade de três bases (TBP, do inglês: *Three-Base Periodicity*) presente nas sequências codificantes [47]. Essa propriedade revela picos na frequência $1/3$ rad/amostra para sequências codificantes de proteínas, enquanto que para sequências não codificantes, esse pico não existe. Por essa razão, a análise no domínio da frequência é apropriada para identificar tal frequência (ou qualquer outra) por meio da análise espectral. É importante destacar que a análise de periodicidades em sequências genômicas no domínio da frequência deve ser feita considerando o sinal resultante da representação numérica do DNA.

3.2.1 Espectro de Energia

A abordagem clássica para análise espectral de uma sequência de DNA foi proposta por David Silverman *et al.* [84]. Os autores definiram e analisaram a transformada de Fourier de uma sequência de bases considerando um mapeamento multidimensional. Nesse caso, a DFT é calculada para cada componente e, por fim, o espectro de energia da sequência é definido como sendo a soma do valor absoluto do quadrado da DFT de cada componente do sinal. Por exemplo, considerando uma sequência representada numericamente pelo mapeamento por indicador binário, o espectro de energia dessa sequência é dado por:

$$S_{\text{multid}}[k] = \left|X_{\mathtt{A}}[k]\right|^2 + \left|X_{\mathtt{C}}[k]\right|^2 + \left|X_{\mathtt{G}}[k]\right|^2 + \left|X_{\mathtt{T}}[k]\right|^2, \tag{3.8}$$

em que $X_\alpha[k]$ com $\alpha \in \{\mathtt{A},\ \mathtt{C},\ \mathtt{G},\ \mathtt{T}\}$ é a DFT do respectivo indicador binários definido como:

$$\mathcal{F}\left[x_\alpha[n]\right] = X_\alpha[k] = \sum_{n=0}^{N-1} x_\alpha[n] e^{-j\frac{2\pi}{N}nk}, \quad k = 0, 1, \ldots, N-1. \tag{3.9}$$

Algumas características de $S_{\text{multid}}[k]$ são apontadas pelo mesmo autor [84]: (i) invariante ao conjunto de eixos ortogonais escolhidos para projeção da representação numérica; e (ii) invariante à permutação do rótulo das bases.

Outra perspectiva da análise no domínio da frequência é considerando um mapeamento unidimensional $\mathcal{M}$, cujo sinal resultante dado na Equação (3.6) é

$$x[n] = a x_{\mathtt{A}}[n] + c x_{\mathtt{C}}[n] + g x_{\mathtt{G}}[n] + t x_{\mathtt{T}}[n].$$

Nesse caso, o espectro de energia é dado por:

$$S[k] = \left|a X_{\mathtt{A}}[k] + c X_{\mathtt{C}}[k] + g X_{\mathtt{G}}[k] + t X_{\mathtt{T}}[k]\right|^2. \tag{3.10}$$

Dependendo se o mapeamento é real ou complexo, o $x[n]$ é também um sinal de valores reais ou complexos. Assim, o espectro de energia será simétrico ou assimétrico com relação

ao eixo das frequências. Portanto, o espectro unilateral deverá ser calculado adicionando o conteúdo das frequências positivas ao conteúdo das frequências negativas.

Uma forma alternativa para reescrever a Equação (3.10) é a forma vetorial. Nesse caso, considerando X_α^k o coeficiente da DFT do sinal $x_\alpha[n]$ em uma determinada frequência k, define-se o seguinte vetor de dimensão 4:

$$\mathbf{X}_k = \left[X_{\mathtt{A}}^k \;\; X_{\mathtt{C}}^k \;\; X_{\mathtt{G}}^k \;\; X_{\mathtt{T}}^k\right]. \tag{3.11}$$

Sendo assim, uma forma alternativa para o espectro de energia é defini-lo na forma vetorial para cada frequência k como:

$$S_k = \mathbf{w}\mathbf{X}_k^*\mathbf{X}_k\mathbf{w}^T, \quad k = 0, 1, \ldots, N-1, \tag{3.12}$$

em que $\mathbf{w}^T$ é o vetor transposto de $\mathbf{w} = [a \;\; c \;\; g \;\; t]$, $\mathbf{X}_k^*$ é a conjugada transposta de $\mathbf{X}_k$ e S_k é um escalar. Portanto, a forma vetorial do espectro de energia é dado por:

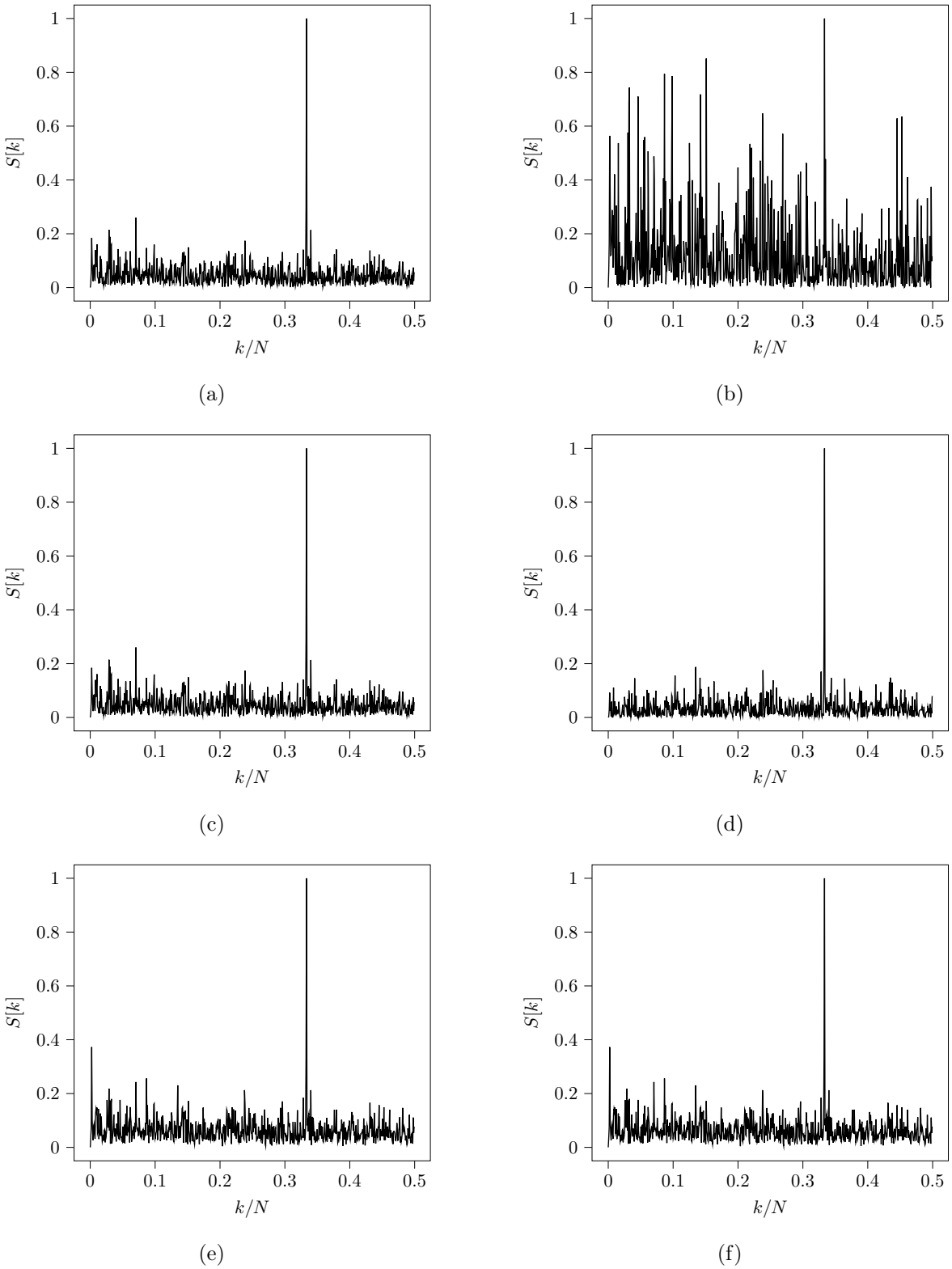
$$\mathbf{S} = [S_0 \;\; S_1 \;\; \cdots \;\; S_{N-1}], \tag{3.13}$$

em que $\mathbf{S}$ é um vetor N-dimensional.

Considerando o gene F56F11.4a que tem sido usado como referência para diferentes técnicas de detecção de regiões exônicas [5, 9, 11, 14], avalia-se o efeito da escolha do mapeamento no cálculo do espectro de energia. Os espectros de energia para a sequência de codificação desse gene ($N = 1350$ bp) considerando os sete mapeamentos apresentados na Tabela 3.1 e o mapeamento adaptativo MEM são apresentados na Figura 3.4. Nesse contexto, o mapeamento CGR foi interpretado como sendo complexo, assim, o sinal resultante foi dado por $x[n] + jy[n]$ [26].

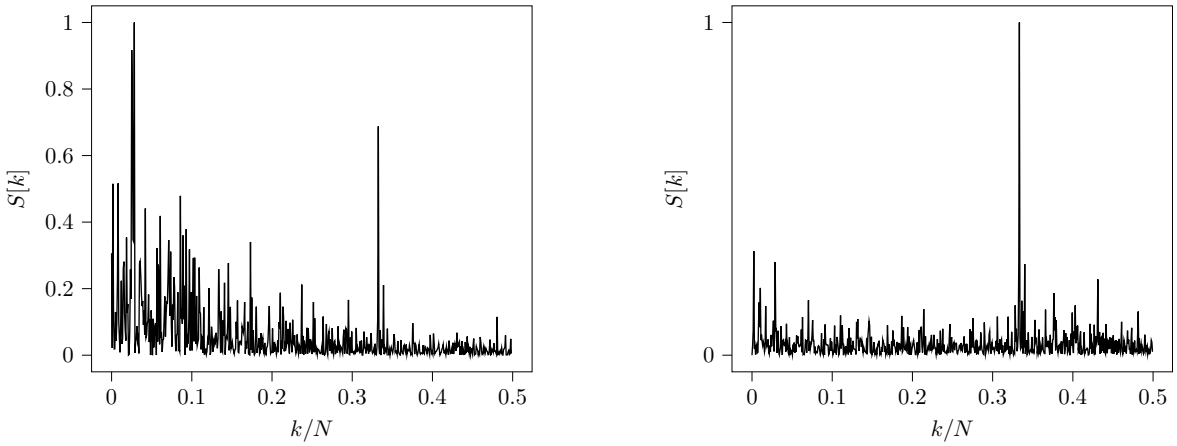
Após analisar a Figura 3.4, destaca-se a importância na escolha do mapeamento adequado para análise espectral. Nem todos os mapeamentos escolhidos revelaram em seu espectro de energia a propriedade TBP que a sequência tem. A existência da propriedade TBP permite que o espectro de uma sequência de DNA atue como um indicador preliminar para uma sequência de codificação revelando um pico na frequência $1/3$ rad/amostra em seu espectro de energia. Para visualizar esse comportamento, considerando o mapeamento indicador binário, o espectro de energia de regiões codificadoras e não codificadoras do mesmo gene são apresentados na Figura 3.5.

Figura 3.4 – Espectro de energia da sequência de codificação do gene F56F11.4a considerando diferentes mapeamentos: (a) Inteiro; (b) Real; (c) Complexo; (d) EIIP; (e) Voss; (f) Tetraedro.



Fonte: Elaborada pela autora

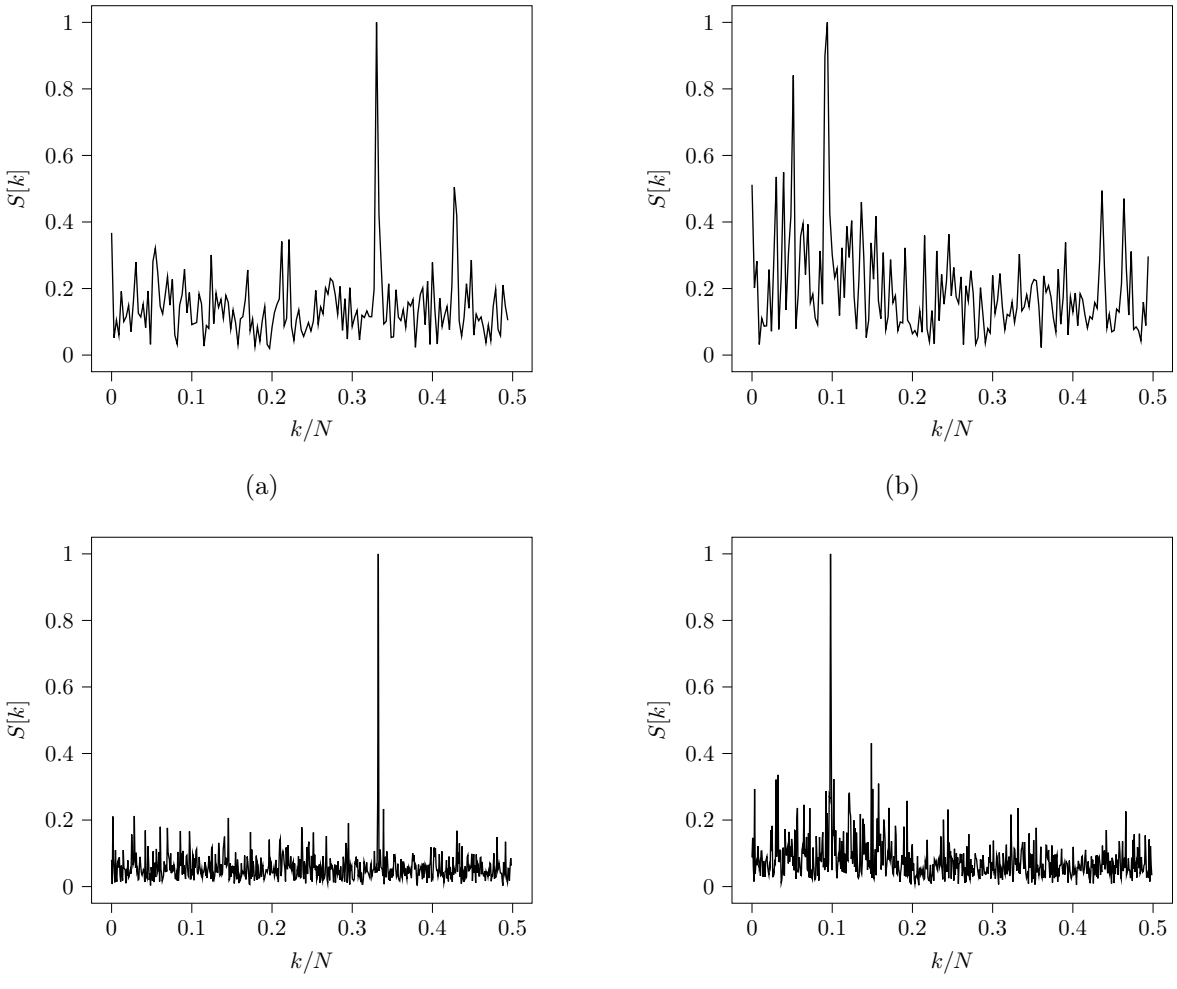
Figura 3.4 – Espectro de energia da sequência de codificação do gene F56F11.4a considerando diferentes mapeamentos: (g) CGR; (h) MEM.



(g) (h)

Fonte: Elaborada pela autora

Figura 3.5 – Espectro de energia para diferentes regiões do gene F56F11.4a: (a) Éxon com 330 bp; (b) Íntron com 330 bp; (c) CDS com 1236 bp; (d) Íntron com 1236 bp.



(a) (b)

(c) (d)

Fonte: Elaborada pela autora

Domínio Tempo-Frequência

Shrish Tiwari *et al.* [4] propuseram um método no qual é suficiente avaliar a energia na frequência 1/3 rad/amostra em janelas de W amostras, deslizando-a por uma ou mais amostras em um processo que analisa toda a sequência de DNA. O espectro de energia é, então, avaliado, também, em função do tempo.

Para isso, uma janela é centrada em diferentes posições ao longo da sequência, e, para cada posição, o espectro de energia é calculado apenas para a frequência 1/3 rad/amostra. Os picos evidenciados nesse espectro ao longo do tempo, devem, portanto, corresponder às regiões exônicas. Os éxons são segmentos do DNA que se unem para forma uma sequência codificantes. Recomenda-se a leitura do Apêndice A para mais detalhes nos conceitos biológicos.

Um critério importante para essa análise é definir o comprimento da janela, M. Shrish Tiwari [4] *et al.* constataram que em sequências genômicas com poucos éxons de comprimento menor que 300 bp, uma janela retangular de comprimento entre 250 e 400 produz resultados semelhantes e satisfatórios. Caso contrário, os autores indicam janelas mais curtas com $M \approx 150$.

Com o intuito de avaliar o efeito da escolha do comprimento da janela para o gene F56F11.4a, cujas posições dos éxons estão listadas na Tabela 3.2, o espectro de energia em função da posição relativa para comprimentos de janela retangular de 10 a 600 é apresentado na Figura 3.6(a). Além disso, os efeitos na mudança do tipo e comprimento da janela para discriminação entre éxons e íntrons são apresentados na Figura 3.6(b) em que a janela é retangular e $M = 351$; na Figura 3.6(c) a janela é retangular e $M = 11$; e na Figura 3.6(d) a janela é Blackman e $M = 351$. Embora quatro dos cinco éxons estejam bem discriminados, o primeiro éxon não está bem discriminado e é facilmente confundido com uma região intrônica. Um motivo para tal confusão é o fato de ser um éxon de comprimento curto. Por essa razão, outras técnicas têm sido investigadas na literatura.

Dimitris Anastassiou [5] propôs determinar um esquema ótimo de mapeamento maximizando a capacidade discriminatória entre genes e sequências pseudo-aleatórias (geradas por meio de um gerador de números aleatórios de quatro símbolos) observando o valor médio e o desvio padrão da DFT de tais sequências. Nesse caso, fixando $c = 0$, as demais variáveis são determinadas maximizando a função:

$$p(a,t,g) = \frac{\mathrm{E}\left\{\left|aX_{\mathtt{A}}\left[N/3\right] + gX_{\mathtt{G}}\left[N/3\right] + tX_{\mathtt{T}}\left[N/3\right]\right|\right\} - \mathrm{E}\left\{\left|aX_{\mathtt{A}_R}\left[N/3\right] + gX_{\mathtt{G}_R}\left[N/3\right] + tX_{\mathtt{T}_R}\left[N/3\right]\right|\right\}}{\mathrm{std}\{|aX_{\mathtt{A}}\left[N/3\right] + gX_{\mathtt{G}}\left[N/3\right] + tX_{\mathtt{T}}\left[N/3\right]|\} + \mathrm{std}\{|aX_{\mathtt{A}_R}\left[N/3\right] + gX_{\mathtt{G}_R}\left[N/3\right] + tX_{\mathtt{T}_R}\left[N/3\right]|\}}, \tag{3.14}$$

para as seguintes condições:

$$\mathrm{E}\left\{\measuredangle\left\{aX_{\mathtt{A}}\left[N/3\right] + gX_{\mathtt{G}}\left[N/3\right] + tX_{\mathtt{T}}\left[N/3\right]\right\}\right\} = 0 \qquad \text{e} \qquad |a| + |g| + |t| = 1,$$

em que $X_{\mathtt{A}_R}$, $X_{\mathtt{G}_R}$, $X_{\mathtt{A}_R}$ são as DFT de cada base da sequência de DNA gerada aleatoriamente por meio de um gerador de números aleatórios de quatro símbolos.

Tabela 3.2 – Localização dos éxons do gene F56F11.4a de *Caenorhabditis elegans*.

Éxon	**Posição Relativa**		**Comprimento**
	Início	**Fim**	**(bp)**
1	928	1039	112
2	2528	2857	330
3	4114	4377	264
4	5465	5644	180
5	7265	7605	465

Para o gene F56F11.4a, a seguinte solução foi encontrada: $a = 0,10 + j0,12$, $c = 0$, $g = 0,45 - j0,19$ e $t = -0,30 - j0,20$. Na Figura 3.6(e), apresenta-se o efeito dessa otimização no processo de identificação de regiões codificadoras desse gene considerando a janela retangular com $M = 351$.

Filtragem Digital

O método de estimativa espectral baseado em Fourier para identificação de regiões exônicas pode ser avaliado sob a perspectiva de filtragem digital. Palghat Vaidyanathan *et al.* [55] propuseram o uso do filtro digital *notch* complementar com banda de passagem estreita centrada em $2\pi/3$. Ao aplicar um sinal na entrada desse filtro, espera-se que o sinal de saída tenha alta energia em regiões com periodicidade 3. De forma geral, um sinal $x[n]$ é aplicado ao filtro e a potência do sinal filtrado, $y[n]$, é $Y[n] = |y[n]|^2$. No caso de uma sequência de DNA mapeada para um sinal multidimensional, por exemplo, indicador binário, essa potência é dada por:

$$Y[n] = |y_{\mathtt{A}}[n]|^2 + |y_{\mathtt{C}}[n]|^2 + |y_{\mathtt{G}}[n]|^2 + |y_{\mathtt{T}}[n]|^2 . \tag{3.15}$$

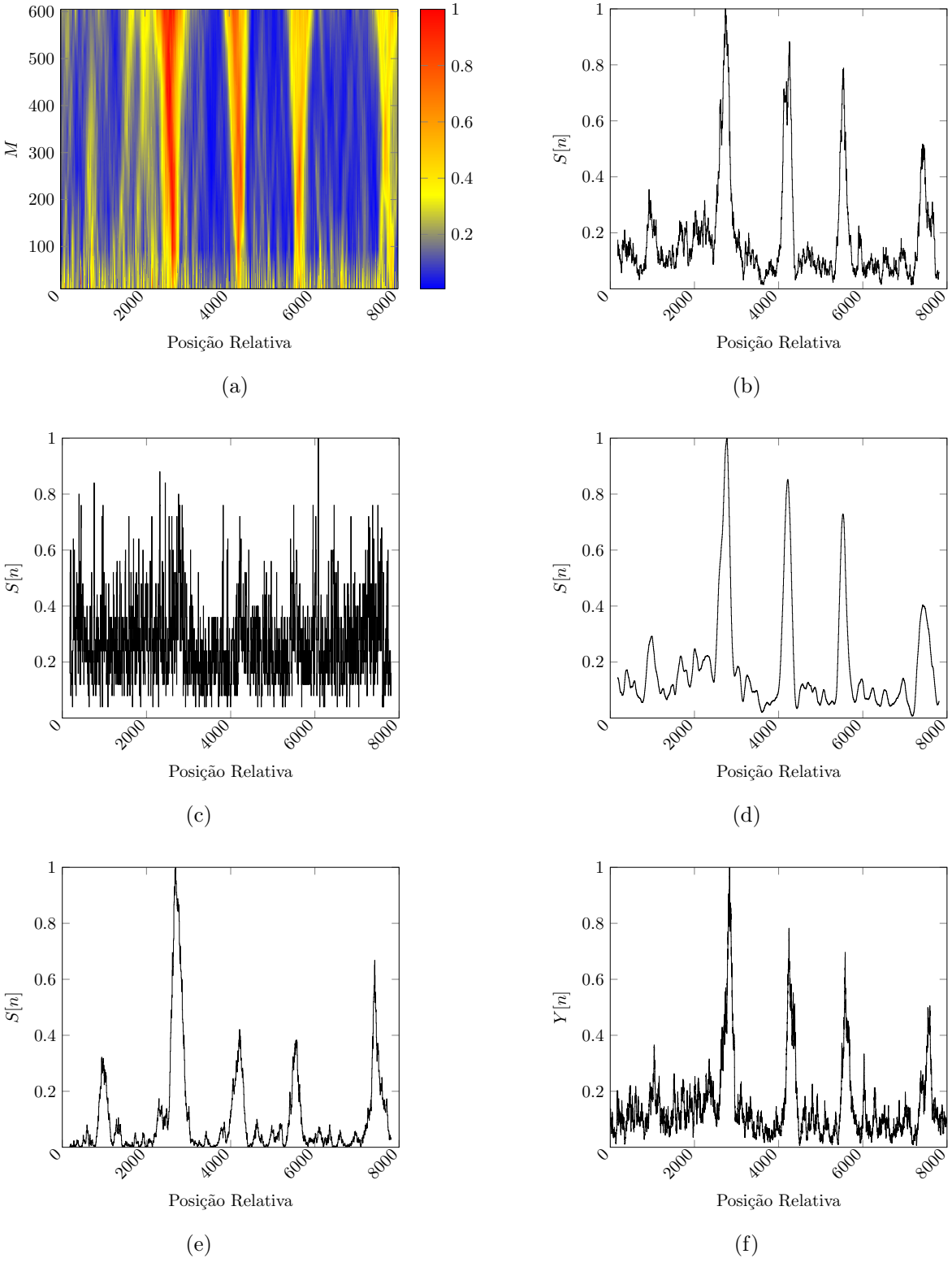
Na Figura 3.6(f) é apresentado o efeito desse processo para identificação das regiões codificadoras do gene F56F11.4a.

3.2.2 Entropia Espectral

A entropia espectral de um sinal é uma medida que caracteriza a não uniformidade da distribuição de energia do sinal no domínio da frequência. O conceito é baseado na entropia de Shannon e interpreta o espectro de energia normalizado no domínio da frequência como uma distribuição de probabilidade [86], ou seja,

$$H(S[k]) = -\sum_{k=0}^{\lfloor N/2 \rfloor} p[k] \log p[k], \tag{3.16}$$

Figura 3.6 – Densidade de energia usando STFT para o gene F56F11.4a considerando: (a) variação no comprimento da janela retangular de 10 a 600; (b) janela retangular e $M = 351$; (c) janela retangular e $M = 11$; (d) janela Blackman e $M = 351$; (e) processo de otimização, janela retangular e $M = 351$; (f) filtro notch complementar com $w_0 = {}^{2\pi}/_{3}$.



Fonte: Elaborada pela autora.

em que,

$$p[k] = \frac{S[k]}{\sum_{k=0}^{\lfloor N/2 \rfloor} S[k]}. \tag{3.17}$$

Nesta tese, o logaritmo natural foi utilizado na Equação (3.16), portanto, a entropia espectral foi dada em nats. Observa-se que a máxima entropia espectral ocorre em um sinal cuja energia está distribuída aproximadamente uniforme nas frequências, nesse caso, $H(S[k]) = \log(N/2 + 1)$ nats. O valor mínimo, $H(S[k]) = 0$, ocorre para um sinal cujo espectro de energia está definido para uma única frequência, ou seja, quando $p[k] = 1$ para algum k.

3.2.3 Envoltória Espectral

A envoltória espectral de um dado sinal é o novo espectro obtido pela maximização do espectro de energia em toda a faixa de frequência $[0, N-1]$. Assim, a envoltória espectral de um sinal $x[n]$ analisado sob a notação vetorial conforme a Equação (3.7),

$$\mathbf{x} = a\mathbf{x}_{\mathtt{A}} + c\mathbf{x}_{\mathtt{C}} + g\mathbf{x}_{\mathtt{G}} + t\mathbf{x}_{\mathtt{T}},$$

é definida em uma determinada frequência k como o espectro máximo sujeito a todos os vetores de peso não triviais possíveis e regularizados por $||\mathbf{w}|| = 1$. Os componentes de $\mathbf{w}$ são números complexos que pertencem a uma hiperesfera complexa de raio unitário.

Em outras palavras, para cada frequência particular k, a envoltória espectral corresponde ao valor máximo de espectro obtido considerando todos os possíveis e não triviais vetores de peso $\mathbf{w}$ tal que a condição $||\mathbf{w}|| = 1$ seja satisfeita. Portanto, após determinar a DFT dos indicadores binários de $\mathbf{x}$, para cada frequência k, a envoltória espectral [57, 56] é dada por:

$$\lambda_k = \max_{\substack{\mathbf{w}\in\mathbb{C}^4 \\ ||\mathbf{w}||=1}} \mathbf{S}_k = \max_{\substack{\mathbf{w}\in\mathbb{C}^4 \\ ||\mathbf{w}||=1}} \mathbf{w}\mathbf{X}_k^*\mathbf{X}_k\mathbf{w}^T. \tag{3.18}$$

É possível notar que essa otimização corresponde a maximização do quociente de Rayleigh para matriz Hermitiana $\mathbf{X}_k^*\mathbf{X}_k$. O quociente de Rayleigh é maximizado quando $\mathbf{w}$ é o autovetor correspondente ao maior autovalor de $\mathbf{X}_k^*\mathbf{X}_k$ (os detalhes da prova estão no Apêndice C). Sendo assim, a decomposição dessa matriz quadrada 4×4 tem a seguinte forma:

$$\mathbf{X}_k^*\mathbf{X}_k = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}, \tag{3.19}$$

em que $\mathbf{Q}$ também é uma matriz quadrada 4×4 cuja i-ésima coluna é o autovetor q_i de $\mathbf{X}_k^*\mathbf{X}_k$, e $\mathbf{\Lambda}$ é uma matriz diagonal cujos elementos são os correspondentes autovalores, $\Lambda_{ii} = \lambda_i$. Sendo assim, a envoltória espectral corresponde ao maior autovalor,

$$\lambda_k = \max_{i\in1,2,3,4} \lambda_i, \tag{3.20}$$

e $\mathbf{w}$ é o seu correspondente autovetor. O pseudocódigo de como determinar λ_k e $\mathbf{w}$ conforme a Equação (3.18) é apresentado no Algoritmo 1.

Algoritmo 1 SpectralEnvelope (s, k)

Entrada: Sequência de DNA s e frequência k
Saída: Mapeamento $\mathcal{M}$
1: Calcular os indicadores binários de s
2: Calcular $\mathbf{X}_k$ usando (3.11)
3: Decomposição de $\mathbf{X}_k^*\mathbf{X}_k \leftarrow \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$
4: $\lambda_k \leftarrow \max(\text{diag}(\mathbf{\Lambda}))$
5: $\mathbf{w} \leftarrow$ autovetor de λ_k
6: **retorna** $\mathbf{w}$

3.3 Complexidade de Sequências de DNA

A análise filogenética de sequências genômicas agrupa informações sobre a diversidade biológica e classificação genética dos organismos, representando, a partir de dendrogramas (ou árvores) a relação evolutiva entre tais organismos. Essa análise é comumente feita utilizando-se de técnicas de alinhamento de sequências. A ideia de alinhamento de sequências tem como base a distância de Levenshtein que avalia o menor número de deleções, inserções e substituições necessárias para transformar uma sequência em outra (detalhes no Apêndice D). Contudo, esses métodos têm um custo computacional alto, especialmente à medida que o comprimento das sequências genômicas crescem.

Portanto, métodos que não necessitem de alinhamento de sequências é de interesse da comunidade científica. A abordagem de agrupamento de dados de expressão gênica provou ser útil para tornar conhecida a estrutura natural inerente aos dados [87]. Porém, a escolha de uma métrica apropriada influencia na formação desses agrupamentos, uma vez que alguns elementos podem estar relativamente mais próximos uns dos outros em uma métrica do que em outra. Nesse contexto, é investigado como a complexidade de sequências de DNA, bem como, a divergência entre representações CGR podem atuar como métricas de dissimilaridade para reconstrução de árvores filogenéticas a partir de sequências de DNA de mamíferos.

A complexidade de Kolmogorov como uma medida de distância entre duas sequências u e v foi inicialmente definida por Chen *et al.* [88] como sendo:

$$d(u,v) = 1 - \frac{K(v) - K(v|u)}{K(uv)}, \tag{3.21}$$

em que $K(\cdot)$ é a complexidade de Kolmogorov, $K(v|u)$ é complexidade condicional de Kolmogorov de v dado u e uv é a sequência resultante da concatenação de u e v. Além disso, $K(v|u)$ é definido como sendo o comprimento do menor programa que faz com que um computador universal padrão produza v dada a entrada u. Assim,

$$K(v|u) = K(vu) - K(u), \tag{3.22}$$

é uma medida que quantifica a aleatoriedade de v dado u.

Para estimar a complexidade de Kolmogorov utiliza-se, por exemplo, a complexidade de Lempel-Ziv [78]. A propriedade da subaditividade da complexidade de Lempel-Ziv garante que, considerando duas sequências u e v, então,

$$\mathcal{C}_{LZ}(uv) \leq \mathcal{C}_{LZ}(u) + \mathcal{C}_{LZ}(v). \tag{3.23}$$

Nesse processo, o grau de similaridade entre u e v é refletido no quanto $\mathcal{C}_{LZ}(uv) - \mathcal{C}_{LZ}(u)$ é menor que $\mathcal{C}_{LZ}(v)$.

***Exemplo* 3.2** *Considerando as sequências* $u =$ `AACGTACCATTG`, $v =$ `CTAGGGACTTAT` *e* $w =$ `ACGGTCACCAA`, *a história exaustiva de cada uma delas é:*

$$\begin{aligned} H(u) &= \texttt{A|AC|G|T|ACC|AT|TG}, \\ H(v) &= \texttt{C|T|A|G|GGA|CTT|AT}, \\ H(w) &= \texttt{A|C|G|GT|CA|CC|AA}, \end{aligned}$$

portanto, $\mathcal{C}_{LZ}(u) = \mathcal{C}_{LZ}(v) = \mathcal{C}_{LZ}(w) = 7$. *Além disso,*

$$\begin{aligned} H(uw) &= \texttt{A|AC|G|T|ACC|AT|TG|ACGG|TC|ACCAA}, \\ H(vw) &= \texttt{C|T|A|G|GGA|CTT|AT|ACG|GT|CA|CC|AA}, \end{aligned}$$

em que $\mathcal{C}_{LZ}(uw) = 10$ *e* $\mathcal{C}_{LZ}(vw) = 12$. *Ou seja, foram necessárias três etapas para construir* w *no processo de produção de* uw*; e cinco etapas para gerar* w *no processo de produção de* vw. *De fato,* u *e* w *compartilham as subsequências* `ACG` *e* `ACC`*; já* v *e* w *compartilham apenas a subsequência* `AC`. *Podemos concluir, portanto, que as sequências* u *e* w *são mais similares que* v *e* w.

Com base nessa ideia de similaridade, alguns autores usaram a complexidade de sequências simbólicas para definir medidas de distância entre as sequências para reconstrução de árvores filogenéticas [89, 23, 24, 90, 91]. Contudo, ao estimar a complexidade de Kolmogorov usando a complexidade de Lempel-Ziv, tem-se que

$$K(v) - K(v|u) \approx K(u) - K(u|v). \tag{3.24}$$

Assim, a distância dada na Equação (3.21) não é exatamente simétrica. Algumas estratégias para torná-la simétrica consiste em considerar o valor mínimo $\min\{K(v) - K(v|u), K(u) - K(u|v)\}$, o valor máximo $\max\{K(v) - K(v|u), K(u) - K(u|v)\}$ ou a média aritmética $(K(v) - K(v|u) + K(u) - K(u|v))/(2)$. A utilização da média aritmética apresentou os melhores resultados em termos de agrupamento e, portanto, foi a escolhida para reproduzir os resultados.

Para obter a árvore filogenética que descreve um conjunto de sequências, primeiramente é determinada a matriz de distância entre essas sequências. Em seguida, essa matriz é utilizada como entrada para o método de grupo de pares não ponderados com

média aritmética (UPGMA, do inglês: *Unweighted Pair Group Method using Arithmetic Mean*) que é um método de clusterização hierárquico aglomerativo usado para gerar as árvores filogenéticas. Indica-se o Apêndice E para mais detalhes sobre esse algoritmo.

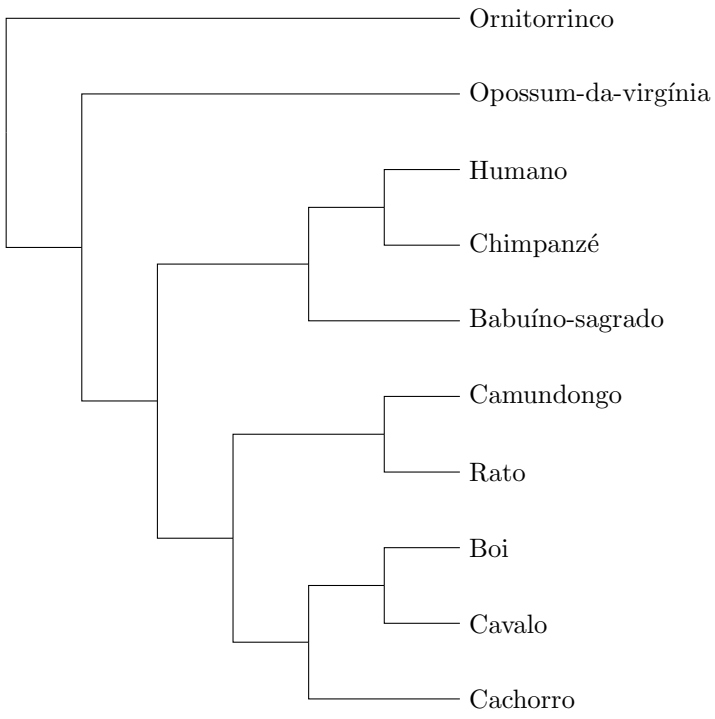
Sendo assim, considerando sequências de DNA mitocondrial de dez espécies de mamíferos placentários (descrição das sequências na Tabela 3.3), obtém-se a árvore filogenética ilustrada na Figura 3.7(a) ao considerar a complexidade de Lempel-Ziv como medida de similaridade entre as sequências. Além disso, investigou-se a performance do algoritmo SEQUITUR para estimação da complexidade de Kolmogorov para esta aplicação, cujo resultado foi publicado em [63] e é apresentado na Figura 3.7(b). O SEQUITUR [92] é um algoritmo de compressão que apresenta complexidade temporal linear [93] e que realiza inferência da estrutura hierárquica de sequências a partir de repetições de subsequências. As repetições encontradas são utilizadas para definir uma gramática em que toda regra deve ser utilizada pelo menos duas vezes e todo par de símbolos terminais ou não-terminais consecutivos que aparecem mais de uma vez devem se tornar regras.

Além dessas abordagens para reconstrução filogenética, as medidas de informação também já foram investigadas como medida de dissimilaridade entre sequências ao considerar suas representações CGR. Para esta aplicação, aproxima-se o CGR por uma imagem com 128 x 128 pixels, na qual, o valor de cada pixel corresponde a quantidade de pontos existentes naquela região normalizados pelo comprimento da sequência. Assim, o valor de cada pixel representa uma probabilidade. Portanto, sendo $\mathrm{P}(x)$ a distribuição de probabilidade do CGR referente a sequência u e $\mathrm{Q}(x)$ a distribuição de probabilidade do

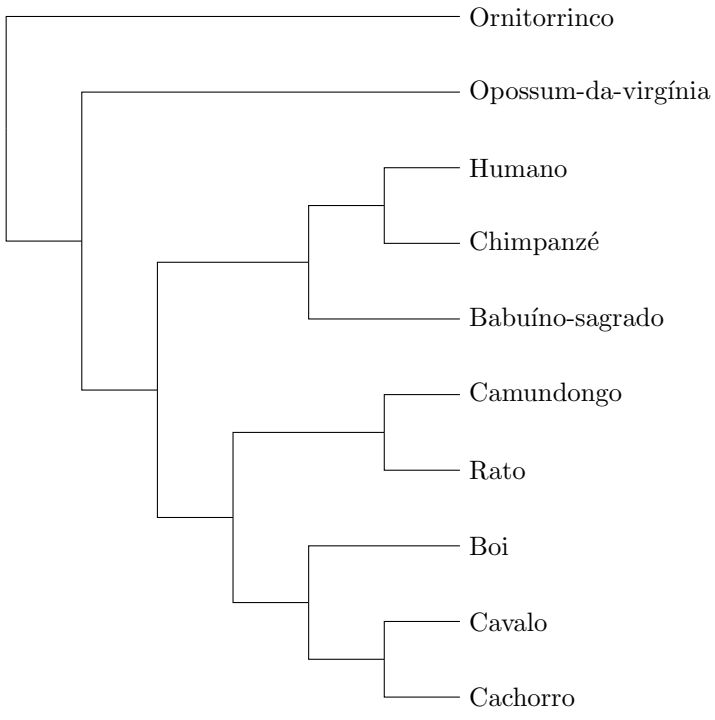
Tabela 3.3 – Especificações das sequências de DNA mitocondrial de dez mamíferos placentários.

Grupo	**Espécie**	**Número GI**	**Comprimento**
Primatas	Chimpanzé	D38116	16554
	Babuíno	U20753	16521
	Humano	D38115	16569
Ferungulata	Cachorro	U96639	16727
	Cavalo	X79547	16660
	Boi	X99256	16338
Roedores	Rato	Y18001	16300
	Camundongo	Y10524	16295
Grupo Externo	Opossum	X14848	17084
	Ornitorrinco	AJ001562	17019

Figura 3.7 – Árvore filogenética obtida para sequências de DNA mitocondrial de dez mamíferos considerando como estimativa da complexidade de Kolmogorov: a) Lempel-Ziv; b) SEQUITUR; c) CGR; d) NCBI.



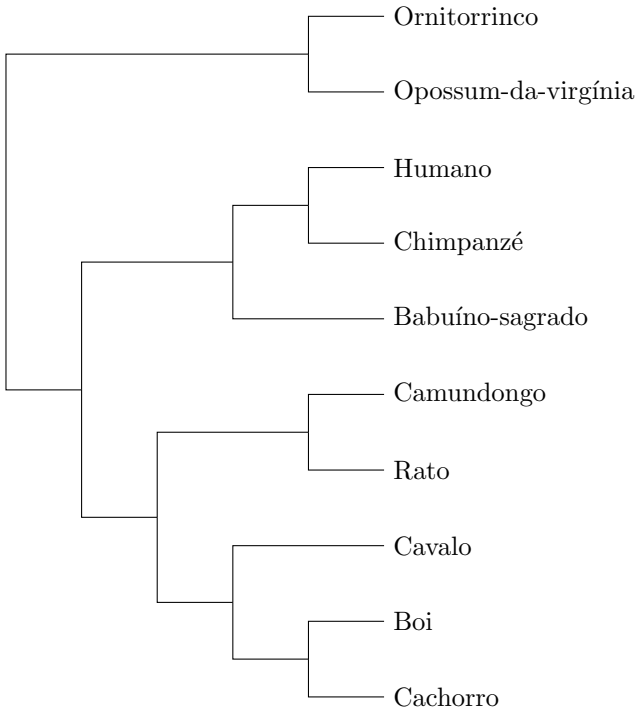
(a)



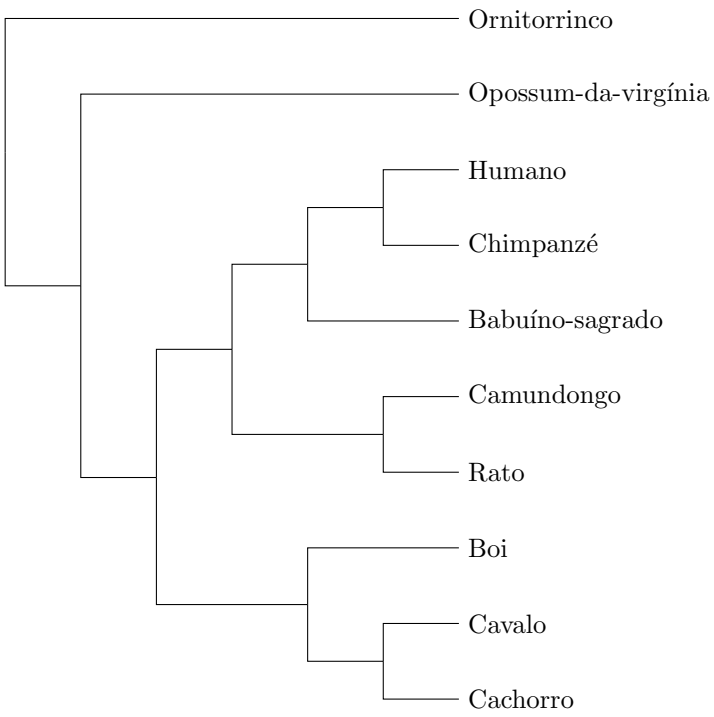
(b)

Fonte: Elaborada pela autora.

Figura 3.7 – Árvore filogenética obtida para sequências de DNA mitocondrial de dez mamíferos considerando como estimativa da complexidade de Kolmogorov: a) Lempel-Ziv; b) SEQUITUR; c) CGR; d) NCBI.



(c)



(d)

Fonte: Elaborada pela autora.

CGR referente a sequência v, a distância entre as sequências é dada por,

$$d(u,v) = \frac{D(\mathrm{P}||\mathrm{Q}) + D(\mathrm{Q}||\mathrm{P})}{2}, \tag{3.25}$$

em que $D(\cdot)$ é a divergência de Kullback-Lieber. A árvore filogenética obtida utilizando esta medida de distância para obtenção de uma nova matriz de distâncias é apresentada na Figura 3.7(c). As representações CGR para as sequências de DNA em consideração podem ser consultadas no Apêndice F.

Embora não exista um consenso sobre toda a árvore filogenética dos mamíferos [94], para esse grupo reduzido de mamíferos espera-se que a árvore tenha o agrupamento apresentado na Figura 3.7(d) conforme referêcia do NCBI [95]. As espécies de primatas, roedores, ferungulata e grupo externo foram associadas corretamente nas árvores de acordo com seus grupos.

As árvores obtidas utilizando tanto o Lempel-Ziv quanto o SEQUITUR quanto o CGR estão próximas do esperado. A única diferença entre as árvores geradas utilizando Lempel-Ziv e SEQUITUR é o agrupamento para o ferungulata, em que o o cachorro e o cavalo são apresentados como mais próximos evolutivamente. Isso pode ter acontecido devido à semelhança entre as sequências do cavalo e cachorro, sendo necessárias sequências maiores para haver uma distinção maior entre as complexidades estimadas. A árvore resultante da análise do CGR foi a mais distante do esperado do NCBI. Especificamente para esse método, o comprimento das sequências é um aspecto ainda mais importante uma vez que ele irá aproximar probabilidades.

O tempo médio para a geração da árvore utilizando o SEQUITUR foi de aproximadamente 1 minuto, enquanto que o tempo médio utilizando o Lempel-Ziv foi de aproximadamente 1 hora e 10 minutos. Isso era esperando uma vez que o SEQUITUR tem complexidade linear no tempo, o que torna possível sua utilização para um maior número de sequências e sequências com maiores comprimentos, ao passo que, sob as mesmas condições, o tempo requerido pelo LZ76 o torna inviável. Já para o CGR o tempo médio foi de 5 segundos. Sendo assim, o CGR pode ser um método atrativo para visualizar agrupamentos gerais, no qual, ainda não é necessário especificar cada um dos sub agrupamentos.

Parte III

Resultados

Capítulo 4

Identificação do DNA a partir de Códigos BCH

As semelhanças entre os sistemas de comunicação biológica e digital têm sido investigadas e novas interpretações conciliando tais teorias têm sido discutidas. Por exemplo, a informação genética de um organismo é codificada em moléculas de DNA por unidades chamadas bases; portanto, pode-se interpretar o conteúdo do DNA como sendo um código digital que representa e transmite a informação armazenada. No entanto, não há um modelo definitivo e a questão de qual código de correção de erros está subjacente às sequências de DNA permanece um problema em aberto. Trabalhos recentes mostraram que as sequências de DNA podem ser identificadas como palavras-código de códigos BCH, uma classe de códigos corretores de erros.

Neste capítulo, propomos melhorias para o algoritmo *DNA Sequence Generation*, principalmente, ao que concerne a construção e decodificação dos códigos. Essas melhorias resultaram em um novo algoritmo para busca de códigos BCH cuja palavra-código difere de uma determinada sequência de DNA (representada numericamente por um mapeamento unidimensional cuja imagem são os elementos do corpo finito $\mathbb{F}_4$) em até um único símbolo. No algoritmo proposto, a decodificação por força bruta é substituída pela decodificação por síndrome.

Com base nesse novo algoritmo, analisa-se, portanto, se em uma coleção de sequências de DNA com a mesma classificação taxonômica existe um código que identifica a maioria dessas sequências, denominado código dominante. Além disso, verifica-se a possibilidade do código dominante fornecer uma informação biológica para a classificação do DNA sendo, assim, um método sem a necessidade de alinhamento de sequências. Por fim, é mostrado que a probabilidade de uma sequência de DNA com comprimento ímpar n ser identificada por um código BCH tende à probabilidade analítica do mesmo código identificar um vetor aleatório. Os resultados desse capítulo foram publicados em periódico [64]. Além disso, os *scripts* estão disponíveis no seguinte repositório do GitHub [96].

4.1 Visão Geral do Algoritmo DNA Sequence Generation

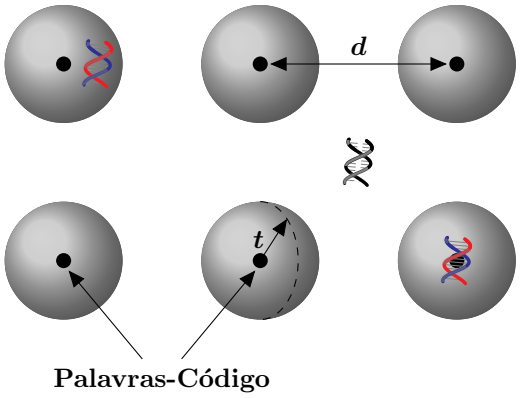
O algoritmo *DNA Sequence Generation* foi proposto por Luzinete Faria *et al.* [42] e Andrea Rocha *et al.* [43] com o propósito de verificar se uma dada sequência de DNA pode ser identificada por um código de correção de erros; mais especificamente, por um código BCH com distância de Hamming $d = 3$. Os autores justificaram que o código BCH foi escolhido devido à simplicidade de seus processos de codificação e decodificação.

Neste cenário, diz-se que uma sequência de DNA é identificada por um código BCH se tal sequência pode ser mapeada para um vetor que é palavra-código de um código BCH ou difere de alguma palavra-código em até um símbolo. Portanto, as três regiões, no espaço vetorial, nas quais o vetor que representa a sequência de DNA pode existir são ilustradas na Figura 4.1. A primeira região é quando esse vetor corresponde exatamente a uma palavra-código; a segunda é a região sombreada, ou seja, quando o vetor está na região de decodificação do código; e a terceira região consiste da região cujos vetores não são decodificáveis pelo código.

Para verificar em qual região está localizada a sequência de DNA, alguns atributos das mesmas devem ser associados aos parâmetros do código. Primeiramente, os códigos BCH são definidos sobre uma estrutura algébrica, portanto, as bases do DNA devem ser mapeadas em conformidade com essa estrutura. As estruturas algébricas utilizadas no *DNA Sequence Generation* foram: corpo finito [42] e anéis de inteiros [43]. Em ambos os casos, as estruturas possuem quatro elementos. Assim, existem vinte e quatro permutações possíveis para mapeamentos unidimensionais bijetivos; porém, de acordo com [42], qualquer um desses mapeamentos apresenta resultados de identificação similares ao utilizar-se códigos BCH sobre $\mathbb{F}_4$.

Um outro atributo importante na busca pela relação entre sequências de DNA e códigos BCH é o comprimento do código. Os autores propuseram, portanto, que a busca por todos os polinômios geradores de códigos BCH de comprimento fixo n, deveria ser feita mediante a construção dos corpos de extensão usando diferentes polinômios primitivos. Porém, este

Figura 4.1 – Regiões de decodificação.



Fonte: Elaborada pela autora.

procedimento resulta em um custo computacional que poderia ser evitado, pois, corpos de extensão isomorfos são construídos repetidamente para cada polinômio primitivo.

Definindo um código BCH cujo comprimento n seja correspondente ao comprimento de uma dada sequência da DNA, para verificar se tal código é capaz de identificar o vetor correspondente a essa sequência, a seguinte verificação é realizada. Primeiramente, a matriz de verificação de paridade do código é usada para decidir se esse vetor é uma palavra-código. Então, o processo de decodificação segue por força bruta. Nesse caso, para cada símbolo do vetor, é testado se ao substituir o símbolo original por algum dos demais três símbolos, esse vetor é validado como palavra-código pela matriz de paridade. Esse processo é, portanto, repetido $3n + 1$ vezes.

Enquanto este algoritmo verifica se uma sequência de DNA é uma palavra-código de um código BCH usando a matriz de verificação de paridade, um novo algoritmo, proposto por Rodríguez *et al.*, usa o método de fatoração polinomial. No entanto, o mesmo processo de força bruta é utilizada para avaliar sequências que diferem em até um símbolo da palavra-código.

4.2 Algoritmo Proposto

O algoritmo começa associando os atributos (alfabeto e comprimento) de sequências de DNA com parâmetros de códigos BCH. O alfabeto do código genético (dado pelo conjunto $\mathcal{N} = \{\mathtt{A},\ \mathtt{C},\ \mathtt{G},\ \mathtt{T}\}$) e o alfabeto de $\mathcal{C}_{BCH}$ (dado pelo conjunto $\mathbb{F}_4 = \{0, \beta, \beta^2, 1\}$) estão relacionados por um mapeamento unidimensional dado por:

$$\mathcal{M} : \mathtt{A} \mapsto 0,\ \mathtt{C} \mapsto \beta,\ \mathtt{G} \mapsto \beta^2,\ \mathtt{T} \mapsto 1, \tag{4.1}$$

conforme mostrado na Figura 4.2. Embora existam vinte e quatro permutações com elementos em $\mathbb{F}_4$, os corpos finitos com o mesmo número de elementos são isomorfos; portanto, consideramos apenas um mapeamento bijetivo. Esta é uma possível razão pela qual Faria *et al.* [42] constata que qualquer mapeamento apresenta resultados de identificação similares, uma vez que, a menos de isomorfismo, existe um único corpo finito com q elementos.

***Exemplo* 4.1** *Considerando que uma sequência de DNA é dada por:*

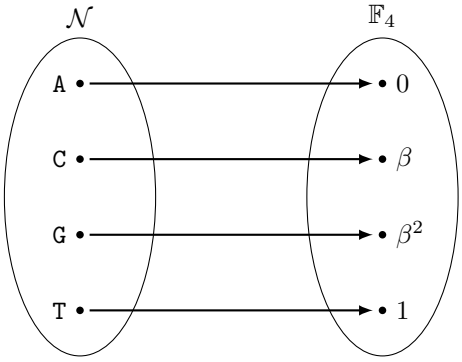
$$s = \mathtt{CTGATCCTTCAAGCG}, \tag{4.2}$$

então, utilizando-se o mapeamento definido na Equação (4.1) têm-se que o vetor resultante é dado por:

$$\mathbf{r} = [\beta \ \ 1 \ \ \beta^2 \ \ 0 \ \ 1 \ \ \beta \ \ \beta \ \ 1 \ \ 1 \ \ \beta \ \ 0 \ \ 0 \ \ \beta^2 \ \ \beta \ \ \beta^2]. \tag{4.3}$$

Uma vez que comprimento da palavra-código n deve corresponder ao comprimento da sequência de DNA, consideraremos apenas as sequências cujo comprimento pode ser

Figura 4.2 – Mapeamento unidimensional dos alfabetos. Cada elemento do código genético, o conjunto $\mathcal{N}$, é mapeado para exatamente um elemento do corpo finito com ordem quatro denotado por $\mathbb{F}_4$.



Fonte: Elaborada pela autora.

escrito como um divisor de $4^m - 1$, em que m é um inteiro positivo. Assim, n é sempre ímpar e satisfaz,

$$n \mid (4^m - 1). \tag{4.4}$$

Portanto, de agora em diante denota-se por $\mathcal{C}_{BCH}$ um código BCH sobre $\mathbb{F}_4$ com parâmetros $[n, k, 3]$.

Na definição formal dos códigos BCH, os diferentes polinômios geradores para códigos $\mathcal{C}_{BCH}$ são especificados por b e ℓ (conforme a Equação (2.10)). Esta foi o primeiro ponto observado como sendo uma oportunidade de melhoria para o *DNA Sequence Generation*. A partir dessa definição, não é mais necessário usar polinômios primitivos diferentes para construir corpos de extensão isomorfos e, assim, determinar todos polinômios geradores de $\mathcal{C}_{BCH}$.

No contexto da decodificação, identificou-se como oportunidade de melhoria a substituição da decodificação por força bruta por decodificação por síndrome. Neste caso, considerando um código $\mathcal{C}_{BCH}$ completamente especificado por um b e um ℓ, verifica-se se o vetor correspondente à sequência de DNA é uma palavra-código ou difere de uma palavra-código em até um símbolo apenas usando o decodificador PGZ, ou seja, resolvendo o sistema não linear da Equação (2.12).

Portando, propõe-se o seguinte algoritmo. Inicia-se o algoritmo com uma sequência de DNA em sua forma vetorial, denotada por $\mathbf{r}$, e um código $\mathcal{C}_{BCH}$ (sugere-se $b = 0$ e $\ell = 1$). Em seguida, verifica-se se ao decodificar $\mathbf{r}$ usando o decodificador PGZ, uma palavra-código válida é retornada; em caso positivo, esse $\mathcal{C}_{BCH}$, representado pelo seu polinômio gerador é salvo em um conjunto $\mathcal{R}$. O algoritmo segue investigando todos os $\mathcal{C}_{BCH}$ cujos polinômios geradores são dados por:

$$g(x) = \text{mmc}(f_b(x), f_{b+\ell}(x)), \tag{4.5}$$

no intervalo $0 \leq b, \ell < n$ tal que $\text{mdc}(n, \ell) = 1$. O algoritmo é finalizado quando todo o

intervalo de b e ℓ é verificado e retorna o conjunto $\mathcal{R}$ de todos os polinômios geradores que identificam a sequência de DNA de entrada. O pseudocódigo é mostrado no Algoritmo 2.

Além disso, duas restrições no intervalo de ℓ reduzem a quantidade de iterações do algoritmo. Primeiro, deve-se considerar apenas um valor para ℓ em cada conjunto $\{s, sq, \ldots, sq^{m-1}\} \mod n$ em que, $0 \leq s < n$. Segundo, deve-se considerar ℓ ou $-\ell$ mod n. Para comprovar a primeira restrição, deve-se lembrar que $\alpha^{q^m} = \alpha$. Assim, considerando os seguintes polinômios geradores,

$$\begin{cases} g_1(x) = \text{mmc}(f_b(x), f_{b+s}(x)), \\ g_2(x) = \text{mmc}(f_{b'}(x), f_{b'+sq}(x)), \\ \qquad \vdots \\ g_j(x) = \text{mmc}(f_{b''}(x), f_{b''+sq^{m-1}}(x)), \end{cases} \tag{4.6}$$

a igualdade $g_1(x) = g_2(x)$ ocorre se $b' = bq$. Da mesma forma, a igualdade $g_1(x) = g_j(x)$ ocorre se $b'' = bq^{m-1}$. Portanto, uma vez que $0 \leq b < n$, é redundante considerar mais de um ℓ em cada conjunto $\{s, sq, \ldots, sq^{m-1}\} \mod n$. Dessa forma, a primeira restrição foi comprovada e a segunda segue o mesmo argumento.

***Exemplo* 4.2** *Considerando um $\mathcal{C}_{BCH}$ de comprimento 15, o intervalo de ℓ deve satisfazer $mdc(\ell, n) = 1$. Assim, $\ell \in \{1, 2, 4, 7, 8, 11, 13, 14\}$. No entanto, a partir da primeira restrição, têm-se que: $\ell \in \{1, 2, 7, 11\}$. Após a segunda restrição, esse intervalo reduz-se a: $\ell \in \{1, 2\}$. Isso reduz consideravelmente o intervalo de ℓ, e, portanto, o custo computacional do algoritmo proposto.*

Uma vez que todos os códigos BCH para uma dada sequência de DNA são conhecidos, o código dominante é encontrado como segue. Considerando $\mathcal{A} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$ uma coleção com N sequências de DNA e $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N$ são os conjuntos de $\mathcal{C}_{BCH}$ que

Algoritmo 2 Algoritmo Proposto

Entrada: Sequência de DNA
Saída: Códigos BCH

```
Inicialize r para Sequência de DNA
Inicialize m
Inicialize range_b
Inicialize range_ℓ
para cada b em range_b faça
    para cada ℓ em range_ℓ faça
        se b e ℓ resolve (2.12) então
            Add g(x) = lcm(f_b(x), f_{b+ℓ}(x)) em R
        fim se
    fim para
fim para
retorna R
```

identificam cada uma delas. O código dominante é o $\mathcal{C}_{BCH}$ que identifica a maioria das sequências de DNA em $\mathcal{A}$. Este código encontra-se na interseção de M conjuntos de $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N$ em que M é o maior número inteiro tal que esta interseção é diferente de zero. Finalmente, $^M/_N$ é a fração das sequências de DNA identificadas pelo mesmo $\mathcal{C}_{BCH}$, o código dominante.

4.3 Resultados

Dados Experimentais

Uma sequência de DNA é referenciada pelo seu número GI. No caso de coleções de sequências de DNA, a seguinte pesquisa é feita no NCBI: txid$X$[Organism] AND "cds"[Feature key] AND Y [Sequence Length] NOT hypothetical protein NOT predicted NOT partial, em que, txid$X$ é substituído pelo identificador taxonômico de um organismo específico (se *Bacteria* txid2, se *Fungi* txid4751, se *Plantae* txid33090) e Y é substituído pelo comprimento de sequência desejada. Além disso, na existência de alguma base degenerada em uma sequência de DNA, a sequência era descartada (por exemplo, `W` é uma base degenerada porque pode representar `A` ou `T`).

Detalhes do Algoritmo

Com algum esforço, os cálculos do algoritmo podem ser feitos sem o auxílio de um *software* computacional para um exemplo simples. Por exemplo, considerando a sequência de DNA 33079225 dada por `CTGATCCTTCAAGCG`. Esta sequência tem comprimento 15 bp. O comprimento $n = 15$ é válido pois pode ser escrito como $4^2 - 1$. Portanto, busca-se por códigos BCH também de comprimento $n = 15$ e cujo grau de extensão do corpo de extensão é $m = 2$.

O algoritmo deve verificar quais códigos BCH sobre $\mathbb{F}_4$ com comprimento $n = 15$ e distância $d = 3$ são capazes de decodificar essa sequência corretamente. Seguindo as etapas do algoritmo proposto, esta sequência é mapeada para:

$$\mathbf{r} = [\beta \;\; 1 \;\; \beta^2 \;\; 0 \;\; 1 \;\; \beta \;\; \beta \;\; 1 \;\; 1 \;\; \beta \;\; 0 \;\; 0 \;\; \beta^2 \;\; \beta \;\; \beta^2], \tag{4.7}$$

cuja representação polinomial é:

$$r(x) = \beta^2 x^{14} + \beta x^{13} + \beta^2 x^{12} + \beta x^9 + x^8 + x^7 + \beta x^6 + \beta x^5 + x^4 + \beta^2 x^2 + x + \beta. \tag{4.8}$$

Inicia-se a especificação do $\mathcal{C}_{BCH}$ por $b = 0$ e $\ell = 1$, ou seja, $g(x) = x^3 + \beta^2 x + \beta$. A solução do decodificador PGZ para esse vetor foi apresentada no Exemplo 2.3. Nesse caso, têm-se que as síndromes são: $s_0 = 1$ e $s_1 = \alpha^{12}$, portanto, sabe-se que o símbolo que diverge da palavra-código ocorreu na décima segunda componente (pois $\alpha^i = \frac{s_1}{s_0} = \alpha^{12}$) e a magnitude desse erro é $\epsilon = 1$. Então, existe uma palavra-código que difere em apenas

Tabela 4.1 – Polinômios geradores no conjunto $\mathcal{R}$ que identificam a sequência de DNA $s = \texttt{CTGATCCTTCAAGCG}$.

b	ℓ	Polinômio Gerador	Posição i	Símbolo	
				Antigo	Novo
0	1	$x^3 + \beta^2 x + \beta$	12	G	C
3	1	$x^4 + \beta x^3 + \beta$	2	G	A
4	1	$x^3 + \beta^2 x^2 + \beta^2$	2	G	A
7	1	$x^4 + \beta^2 x^3 + \beta^2 x^2 + \beta^2 x + 1$	10	A	G
10	1	$x^3 + x + \beta$	3	A	C
14	1	$x^3 + \beta x^2 + \beta^2$	13	C	G
0	2	$x^3 + \beta x + \beta^2$	0	C	G
3	2	$x^3 + x^2 + \beta$	2	G	A
5	2	$x^3 + x + \beta^2$	0	C	A
8	2	$x^3 + \beta x^2 + \beta$	10	A	G
10	2	$x^3 + \beta^2 x + \beta^2$	7	T	C
13	2	$x^3 + \beta^2 x^2 + \beta$	5	C	G

um símbolo do vetor representativo da sequência da DNA. Essa palavra-código é dada por:

$$\mathbf{c} = [\beta \;\; 1 \;\; \beta^2 \;\; 0 \;\; 1 \;\; \beta \;\; \beta \;\; 1 \;\; 1 \;\; \beta \;\; 0 \;\; 0 \;\; \beta \;\; \beta \;\; \beta^2], \tag{4.9}$$

e corresponde à sequência CTGATCCTTCAACCG. Este polinômio gerador é, por fim, salvo em $\mathcal{R}$.

Essa operação é repetida para todos os demais códigos especificados por $0 \leq b < n$ e $\ell \in \{1, 2\}$. Após analisar todo esse intervalo, obtém-se doze códigos no conjunto $\mathcal{R}$. Esses códigos têm polinômios geradores conforme listado na Tabela 4.1. Além dessa informação, também são listadas a posição em que existe divergência entre a sequência de DNA e a palavra-código, o símbolo antigo e o novo símbolo nesta posição. O símbolo antigo corresponde a base na posição i da sequência de DNA original. O novo símbolo é a base na posição i da palavra-código cuja distância da sequência de DNA é no máximo um.

Análise Biológica

Considerando o exemplo da seção anterior, todos os códigos do conjunto $\mathcal{R}$ que identificaram a sequência de DNA, decodificavam a sequência original para uma palavra-código que diferia em um símbolo. No contexto biológico, essa diferença de um símbolo pode significar um polimorfismo de nucleotídeo único (SNP, do inglês: *Single*

Nucleotide Polymorfism) ou uma sequência ancestral (nesse caso, dada a característica de correção de erro do código BCH, esse pode estar atuando na proteção contra mutações).

Especificamente para o código cujo polinômio gerador é descrito por $g(x) = x^3 + \beta^2 x + \beta$, a diferença entre os símbolos ocorre na posição 12, resultando em uma mutação do tipo transversão. As transversões são um tipo de mutação pontual que altera uma base purina (`A` ou `G`) para uma pirimidina (`C` ou `T`) ou de uma pirimidina para uma purina. Além disso, também identificaram a sequência de DNA por uma transversão os seguintes códigos representados pelo par (b, ℓ): (10,1), (14,1), (0,2), (5,2) e (13,2). Para os demais códigos no conjunto $\mathcal{R}$, o SNP causa uma mutação do tipo transição. As mutações do tipo transição referem-se a alterações de uma base purina para outra purina, ou uma base pirimidina para outra pirimidina.

Embora o SNP represente uma mutação, ele pode ser silencioso durante a tradução de proteínas. Isso quer dizer que, independente da mudança da base, o mesmo aminoácido continua sendo produzido por aquele códon. Este fenômeno acontece, por exemplo, com a sequência de DNA cujo número GI é 1852346641. Trata-se de uma sequência do genoma do organismo *Streptomyces coelicolor* que tem 255 bp. É uma molécula do tipo DNA genômico cujo produto é o transportador de proteínas MFS (uma das duas maiores famílias de transportadores de membrana encontrado na Terra [97]). Esta sequência de nucleotídeos é traduzida para a proteína WP_173944011.1 cuja leitura de códons inicia na posição 3. O algoritmo proposto retornou, para essa sequência, um conjunto $\mathcal{R}$ com 128 códigos que a identificam, entre os quais, destaca-se:

$$\begin{aligned} g_1(x) &= x^5 + \beta^2 x^4 + \beta^2 x^3 + x^2 + 1, \\ g_2(x) &= x^5 + x^4 + \beta x^3 + \beta. \end{aligned} \tag{4.10}$$

Na Tabela 4.2 é apresentada a sequência de DNA original e as palavras-código para os códigos $g_1(x)$ e $g_2(x)$ que estão subjacentes à sequência original devidamente mapeadas em sequências de DNA. As seguintes abreviaturas são úteis: *Ont* é o sequência de DNA original, *Gnt1* é a sequência de bases subjacente à sequência original a partir da palavra-código de $g_1(x)$ e *Gnt2* é a sequência de bases subjacente à sequência original a partir da palavra-código de $g_2(x)$. As bases em que essas sequências diferem estão destacadas em vermelho e os respectivos códons estão sublinhados.

Nesse caso, observa-se que entre *Ont* e *Gnt1*, o SNP resulta em uma mutação do tipo transição (`G` $\rightarrow$ `A` na posição 155); no entanto, o códon no qual ocorreu a incompatibilidade é traduzido para o mesmo aminoácido, um ácido aspártico cuja abreviatura é Asp ou D. No contexto biológico, essa incompatibilidade representa uma mutação silenciosa. Por outro lado, embora entre *Ont* e *Gnt2*, também haja uma mutação do tipo transição (`A` $\rightarrow$ `G` na posição 204), o respectivo códon é traduzido para diferentes aminoácidos, mudando de uma Leucina (Leu ou L) para uma Prolina (Pro ou P). A tradução de códons é feita usando a Tabela A.1. Por se tratar de uma molécula de DNA, essa sequência deve ser

Tabela 4.2 – Sequência do *Streptomyces coelicolor* com número GI 1852346641.

Posição		**Sequência**		
1	*Ont*:	GCTGGGAGAC	GGCGATGCCG	ATGACGATCG
	Gnt1:	GCTGGGAGAC	GGCGATGCCG	ATGACGATCG
	Gnt2:	GCTGGGAGAC	GGCGATGCCG	ATGACGATCG
31	*Ont*:	CGGCCTGCTG	GAAGGAACCG	AGCCGGCCGC
	Gnt1:	CGGCCTGCTG	GAAGGAACCG	AGCCGGCCGC
	Gnt2:	CGGCCTGCTG	GAAGGAACCG	AGCCGGCCGC
61	*Ont*:	GGTAGGCGGG	CGGGGACACC	TCGGCGATGT
	Gnt1:	GGTAGGCGGG	CGGGGACACC	TCGGCGATGT
	Gnt2:	GGTAGGCGGG	CGGGGACACC	TCGGCGATGT
91	*Ont*:	AGGCGGGGCC	GATCACCGAG	GCCATGCCGA
	Gnt1:	AGGCGGGGCC	GATCACCGAG	GCCATGCCGA
	Gnt2:	AGGCGGGGCC	GATCACCGAG	GCCATGCCGA
121	*Ont*:	TCGCGAAACC	GCCGATGATC	CGCCACATGG
	Gnt1:	TCGCGAAACC	GCCGATGATC	CGCCACATGG
	Gnt2:	TCGCGAAACC	GCCGATGATC	CGCCACATGG
151	*Ont*:	CCAGGTCCCA	CAGCGCGAAG	GGCAGCGCGG
	Gnt1:	CCAGATCCCA	CAGCGCGAAG	GGCAGCGCGG
	Gnt2:	CCAGGTCCCA	CAGCGCGAAG	GGCAGCGCGG
181	*Ont*:	AGCCGACGGC	GCTCACCGTG	AACAGGACCG
	Gnt1:	AGCCGACGGC	GCTCACCGTG	AACAGGACCG
	Gnt2:	AGCCGACGGC	GCTCACCGTG	AACGGGACCG
211	*Ont*:	CGGCGATCTG	CATGCAGCGG	ATACGGCCGA
	Gnt1:	CGGCGATCTG	CATGCAGCGG	ATACGGCCGA
	Gnt2:	CGGCGATCTG	CATGCAGCGG	ATACGGCCGA
241	*Ont*:	AACAGGACCG	GATAC	
	Gnt1:	AACAGGACCG	GATAC	
	Gnt2:	AACAGGACCG	GATAC	

convertida em seu complemento reverso antes de traduzir os códons. Entre os 128 códigos que a identificam, 47 resultam em uma mutação silenciosa.

Performance do Algoritmo

Para efeito de comparação, tentou-se considerar o mesmo conjunto de sequências de DNA do trabalho de Luzinete Faria *et al.* [42], porém, alguns registros já foram removidos do banco de dados. Em geral, o algoritmo proposto retornou mais códigos BCH que identificam as sequências de DNA analisadas. Por exemplo, para as sequências 78096542 e 45368559, em vez de apenas um código BCH, o algoritmo proposto retorna 34 e 92 códigos, respectivamente. Já para as sequências 51093376 e 832917, em vez de dois códigos

BCH, o algoritmo proposto retorna 32 códigos para ambos. É importante destacar que, os códigos BCH encontrados por [42] também foram encontrados por nós. Porém, o algoritmo proposto tem encontrado mais códigos que identificam cada sequência de DNA.

Comparado ao algoritmo proposto por Rodríguez-Sarmiento [33], o algoritmo proposto neste capítulo retornou exatamente os mesmos códigos para as seguintes sequências de DNA HP283558.1, HP425961.1, HP347514.1, HP253977.1, HP296666.1, HP320974.1, HP352962.1, HP278326.1, EZ071796.1 e EZ111718.1. Para as demais sequências (AK280992.1, HP466062.1, HP933108.1, HP823668.1), o algoritmo em [33] retorna códigos BCH com $d \geq 3$. Os códigos BCH com $d \geq 3$ são subcódigos de códigos com $d = 3$ e, portanto, existem palavras-código comuns entre eles. Assim, se uma dada sequência é identificada por um código $\mathcal{C}_1$ com $d > 3$ e polinômio gerador $g_1(x)$, então, ela também é identificada por um código $\mathcal{C}_2$ com $d = 3$ e polinômio gerador $g_2(x)$, em que $g_1(x)$ divide $g_2(x)$. Nesse caso, ambos os códigos traduzem a sequência original para a mesma palavra-código.

Por exemplo, considerando a sequência de DNA AK280992.1, o algoritmo proposto retorna $\mathcal{R}$ com cardinalidade 18. Para este exemplo em particular, todos os códigos em $\mathcal{R}$ traduzem a sequência original para a mesma palavra-código que tem 18 raízes consecutivas. Portanto, existe um subcódigo com $d > 3$ completamente definido por esse número de raízes consecutivas, no qual, esta palavra-código também é palavra-código do subcódigo. Sendo assim, essa sequência de DNA também é identificada pelo código $[95, 5, 19]$ com o seguinte polinômio gerador,

$$\begin{aligned} g(x) =& x^{90} + x^{85} + x^{80} + x^{75} + x^{70} + x^{65} + x^{60} + x^{55} + x^{50} + x^{45} + x^{40} + \\ & x^{35} + x^{30} + x^{25} + x^{20} + x^{15} + x^{10} + x^5 + 1. \end{aligned} \tag{4.11}$$

Sendo assim, saber quais os códigos com $d = 3$ identificam uma determinada sequência de DNA é suficiente para encontrar subcódigos com $d > 3$. Para isso, realiza-se o processo de fatoração uma única vez. As melhorias no processo de decodificação, substituindo a força bruta pelo decodificador PGZ, resultam em menos operações realizadas. Consequentemente, o tempo de execução do algoritmo proposto é menor.

Análise Estatística

A probabilidade de um $\mathcal{C}_{BCH}$ identificar uma sequência de DNA pode ser interpretada como a probabilidade de uma sequência de DNA estar no raio de decodificação do código. As três regiões do espaço vetorial que uma sequência de DNA pode ocupar já foram discutidas e apresentadas na Figura 4.1. Em geral, o cálculo da probabilidade desses eventos não é trivial, mas uma probabilidade analítica satisfatória pode ser calculada em alguns casos especiais [69].

Para isso, faz-se uma analogia entre problema de decodificação da sequência de DNA e o de decodificação usando códigos lineares para canais que cometem erros de símbolo

independentemente. Nesse último caso, um decodificador decodificará cada vetor recebido para a palavra-código mais próxima, desde que esteja a uma distância t da palavra-código. Podemos analisar o desempenho desse decodificador quando um vetor cujos símbolos são independentes e identicamente distribuídos (iid) é selecionado do espaço vetorial n-dimensional sobre $\mathbb{F}_4$. A probabilidade desse vetor ser identificado por um código linear (como o código BCH) é a probabilidade de um vetor ser decodificado corretamente.

Assim, especificamente para os códigos com $d = 3$, a probabilidade de decodificação correta é dada pela a soma de q^k palavras-código e $n(q-1)$ palavras cuja distância de Hamming da palavra-código é unitária dividida pelo total de palavras em um espaço vetorial n-dimensional, q^n, ou seja,

$$\mathrm{P} = \frac{q^k + q^k n(q-1)}{q^n}, \tag{4.12}$$

ou ainda,

$$\mathrm{P} = \left(\frac{1}{q^{n-k}} + \frac{n(q-1)}{q^{n-k}} \right). \tag{4.13}$$

Essa probabilidade é maximizada quando o código tem a maior dimensão k possível. Nesse caso, o polinômio gerador possui grau mínimo. Assim, dado todos os códigos BCH definidos sobre um espaço vetorial $\mathbb{F}_4$ cujo comprimento é n, essa probabilidade é maximizada para o código cujo grau do polinômio gerador é o menor possível. Isto é, a probabilidade é maximizada quando,

$$n - k = \min_{\mathcal{C}_{BCH}} \deg g(x). \tag{4.14}$$

A probabilidade de que um vetor n-dimensional sobre $\mathbb{F}_4$ seja identificado por um $\mathcal{C}_{BCH}$ com polinômio gerador de grau mínimo é, portanto, uma função de n. Essa probabilidade foi calculada para alguns códigos e é apesentada na Tabela 4.3. Observa-se que alguns códigos têm probabilidade igual a um, são os códigos perfeitos. Um código perfeito é aquele para o qual existem esferas de raio igual em torno das palavras de código que são disjuntas e que preenchem completamente o espaço [69]. Verifica-se, portanto, que os códigos $[341, 336, 3]$ e $[5461, 5454, 3]$ são códigos perfeitos.

Tabela 4.3 – Probabilidade de que um vetor n-dimensional sobre $\mathbb{F}_4$ cujos símbolos são iid seja identificado por um $\mathcal{C}_{BCH}$ no qual o polinômio gerador tem grau mínimo.

$\boldsymbol{m}$	$\boldsymbol{n}$	$\boldsymbol{\min_{\mathcal{C}_{BCH}} \deg g(x)}$	$\mathbf{P}$
4	255	5	0.75
5	341	5	1
6	273	7	0.05
7	5461	7	1
12	7735	10	0.022

Tabela 4.4 – Códigos dominantes para diferentes coleções de sequências de DNA.

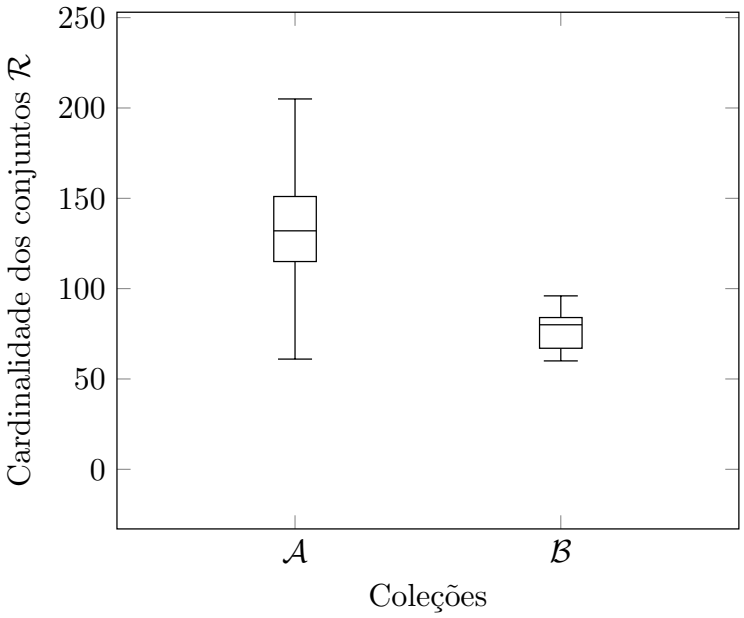
n	**Kingdom**	N	$\frac{M}{N}$	$g(x)$	**P**
255	Bacteria	500	77,15 %	$x^5 + x^4 + \beta^2 x^3 + x + \beta$	75%
	Bacteria	700	75,86 %	$x^5 + \beta x^3 + \beta x^2 + x + \beta$	
	Fungi	500	78,74 %	$x^5 + \beta x^2 + 1$	
	Plantae	500	83,99 %	$x^5 + \beta x^4 + \beta x^3 + x^2 + \beta^2 x + 1$	
341	Bacteria	500	100%	$x^5 + \beta^2 x^3 + \beta x^2 + \beta x + 1$	100%
	Fungi	479			
	Plantae	500			
273	Bacteria	500	8,45 %	$x^7 + \beta x^6 + \beta^2 x^5 + \beta^2 x^4 + x^3 + x^2 + \beta x + 1$	5%
	Bacteria	2000	6,37 %	$x^7 + x^6 + x^4 + \beta^2 x^3 + \beta^2 x + 1$	
	Fungi	1800	7,91 %	$x^7 + \beta x^5 + x^4 + x^2 + \beta$	
	Plantae	990	10,04 %	$x^7 + \beta^2 x^6 + x^4 + \beta x^3 + \beta^2 x + \beta^2$	

A probabilidade de um dado $\mathcal{C}_{BCH}$ identificar sequências de DNA pode ser estimada computacionalmente. Para tanto, observou-se em coleções com N sequências de DNA com a mesma classificação taxonômica, qual a fração de sequências que são identificadas pelo mesmo código (o código dominante). Deseja-se verificar se a taxa $^M/_N$ tende a probabilidade P em que M é a quantidade de sequências identificadas pelo código dominante. Em caso positivo, é possível concluir que as sequências de DNA estão distribuídas de maneira aproximadamente uniforme, sob a métrica de Hamming, em um espaço vetorial n-dimensional. Caso contrário, conclui-se que podem existem diferentes códigos dominantes para diferentes coleções, assim, seria possível distinguir e classificar duas ou mais coleções.

Os códigos dominantes para coleções de N sequências de DNA de acordo com a classificação de três reinos especificados no início da seção 4.3 são apresentados na Tabela 4.4. É possível observar que, em geral, a porcentagem de sequências identificadas por um código (ou seja, $^M/_N$) tende à probabilidade analítica do mesmo código identificar um vetor escolhido aleatoriamente. Isso é especialmente observado ao comparar as duas primeiras linhas (ou as linhas oito e nove) da Tabela 4.4, na qual à medida que cardinalidade das coleções cresce, a taxa $^M/_N$ se aproxima ainda mais de P. Observa-se, também, que, conforme esperado, um código perfeito pode identificar qualquer vetor n-dimensional e, consequentemente, qualquer sequência de DNA. Na Tabela 4.4, o código perfeito é o $[341, 336, 3]$. Além disso, todos os códigos dominantes possuem grau mínimo.

Resultados semelhantes foram obtidos mesmo ao analisar coleções cuja classificação taxonômica é menos generalista. Para exemplificar, considera-se a análise de sequências de organismos *Streptomyces* (txid1883), o maior gênero de *Actinobacteria*. Neste caso, duas coleções $\mathcal{A}$ e $\mathcal{B}$ são analisadas cujo comprimento das sequências em cada coleção é 255 bp e 341 bp, respectivamente. Os códigos dominantes retornados pelo algoritmo

Figura 4.3 – *Boxplot* da cardinalidade dos conjuntos $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N$ para cada sequência de DNA em ambas as coleções $\mathcal{A}$ e $\mathcal{B}$.



Fonte: Elaborada pela autora.

proposto são:

$$\begin{aligned} g_{\mathcal{A}}(x) &= x^5 + \beta^2 x^4 + \beta^2 x^3 + x^2 + 1, \\ g_{\mathcal{B}}(x) &= x^5 + \beta^2 x^3 + \beta^2 x^2 + \beta^2 x + 1, \end{aligned} \tag{4.15}$$

em que $g_{\mathcal{A}}(x)$ identifica cerca de 78,27 % da coleção $\mathcal{A}$, e $g_{\mathcal{B}}(x)$ identifica 100 % da coleção $\mathcal{B}$. Mais uma vez, os códigos dominantes têm grau mínimo e a taxa de identificação tende ao valor analítico de P para identificação de um vetor escolhido aleatoriamente.

A dispersão de cardinalidade dos conjuntos $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_N$ para ambas as coleções é apresentada na Figura 4.3 por meio do gráfico do tipo diagra de caixas (ou *boxplot*). O *boxplot* é uma representação estatística de dados que exibe informações sobre a distribuição dos dados por meio de seus quartis para uma ou mais categorias de dados. Para o caso da coleção $\mathcal{A}$, a cardinalidade desses conjuntos tem uma média de $134,85$ códigos com desvio padrão de $79,70$. Já para coleção $\mathcal{B}$, a cardinalidade tem uma média de $77,14$ códigos com padrão desvio de $9,37$ (alguns códigos com $n = 341$ são códigos perfeitos).

A posição da mediana centralizada entre os quartis do *boxplot* para coleção $\mathcal{A}$, revela que a cardinalidade de $\mathcal{R}_i$ é normalmente distribuída. Já para coleção $\mathcal{B}$, a mediana está mais próxima do topo da caixa; então as cardinalidades não estão normalmente distribuídas. Ainda é possível observar para esse caso que existem 60 códigos perfeitos no espaço vetorial no qual as sequências da coleção $\mathcal{B}$ são mapeadas. Por esta razão, a cardinalidade de $\mathcal{R}_i$ para sequências na coleção $\mathcal{B}$ é pelo menos o número de códigos perfeitos com $n = 341$, ou seja, 60.

Em todos os casos anteriores, os códigos dominantes são códigos cujo polinômio gerador tem grau mínimo. Esses códigos identificam as coleções com probabilidade tendendo ao valor analítico de P (probabilidade de identificação de um vetor escolhido aleatoriamente). Este é um indicativo de que os códigos dominantes não podem fornecer classificação biológica.

4.4 Considerações

O algoritmo proposto para identificar sequências de DNA como palavras-código de códigos BCH sobre o corpo finito $\mathbb{F}_4$ foi implementado na linguagem baseada em *Python*, *SageMath*, um sistema de software matemático de código aberto gratuito. O algoritmo proposto usa teoria da informação e abordagens de álgebra abstrata para melhorar a demanda computacional observada nos algoritmos propostos em [42, 33]. Nesse cenário, diz-se que uma sequência de DNA é identificada por um código BCH $[n, k, 3]$ sobre $\mathbb{F}_4$ se essa sequência, ao ser devidamente mapeada para um vetor no espaço n-dimensional do $\mathbb{F}_4$, difere em até um símbolo de alguma palavra-código.

As melhorias propostas para o algoritmo *DNA Sequence Genarion* resultaram em um novo algoritmo, que foi proposto nesse capítulo. Entre as melhorias propostas, considera-se que a mais importante foi substituir a decodificação de força bruta pelo decodificador PGZ, resultando, assim, em uma redução significativa no número de operações para análise de uma sequência. Isso porque, nos algoritmos anteriores, a decodificação era realizada por um processo de localização de raiz [33] ou de verificação da matriz de paridade [42] repetindo esse processo $3n + 1$ vezes (nesse ciclo, as outras três bases do DNA eram testadas para todas as posições da sequência a fim de obter o conjunto de sequências vizinhas). Como o algoritmo proposto neste capítulo propõe o uso do decodificador PGZ, a solução se restringe a resolver a Equação (2.12) para intervalos delimitados de b e ℓ, assim, o número de operações é reduzido. E, consequentemente, reduz-se o tempo de execução do algoritmo. Além disso, em geral, o algoritmo proposto retorna $\mathcal{R}_i$ com cardinalidade maior que um. Isso representa um diferencial em relação ao algoritmo em [42]. Porém, embora a mesma cardinalidade seja esperada como resultado do algoritmo em [33], o tempo de execução do algoritmo proposto é menor.

Além disso, mostrou-se que as sequências de DNA (com comprimento n em conformidade com a Equação (4.4)) estão distribuídas de maneira aproximadamente uniforme, sob a métrica de Hamming, em um espaço vetorial de dimensão n. Os códigos dominantes não fornecem informações biológicas suficientes para coleções de sequências de DNA a fim de permitir a classificação sem a necessidade de alinhamento das sequências. Isto é, as sequências de DNA mapeadas para palavras-código não podem ser classificadas apenas pelos polinômios geradores.

Embora esta não seja uma resposta definitiva para a questão da existência ou não

de um código BCH subjacente às sequências de DNA, verificamos que o SNP na posição apontada pelo código não influencia esta classificação. Verificamos por meio de simulações que não há perda de informação biológica quando mapeamos uma sequência de DNA para uma palavra-código de algum $\mathcal{C}_{BCH}$ e a classificamos usando, por exemplo, o método Kameris [27].

Capítulo 5

Discriminação de Sequências Codificantes

As abordagens de processamento de sinal digital foram investigadas como um indicador preliminar para discriminar entre as sequências codificantes e não codificantes de proteínas do DNA. Isso ocorre porque já foi comprovada a existência de uma periodicidade de três bases em regiões codificadoras de proteínas. Assim, observando o espectro de energia de uma sequência de codificação, essa periodicidade reflete em um pico proeminente no na frequência $\frac{1}{3}$ rad/amostra. Contudo, para esta análise, uma vez que as sequências de DNA são sequências simbólicas, elas devem ser mapeadas em um ou mais sinais, de modo que essa informação seja destacada e a análise espectral possa ser realizada.

Neste capítulo, propomos, portanto, dois novos algoritmos para calcular mapeamentos adaptativos e, por meio desses, discriminar sequências codificantes a partir da observação do respectivo espectro de energia. Ambos os algoritmos são baseados na abordagem de envoltória espectral. Por fim, verificamos o melhor desempenho dos novos métodos considerando tanto sequências de DNA sintéticas quanto reais em comparação aos métodos: Voss [53], EIIP [54], QPSK [5] e o MEM [8] que também é um método adaptativo. Demonstraremos que nosso método tem maiores taxas de acurácia e especificidade na discriminação de sequências codificantes. Isso é especialmente importante nesta aplicação, pois reduz os riscos de uma sequência codificante não ser identificada. Os resultados desse capítulo foram publicados em periódico [61]. Além disso, os *scripts* estão disponíveis no seguinte repositório do GitHub [98].

5.1 Representação Numérica Adaptativa

Em conformidade com os métodos de processamento genômico, o primeiro procedimento para a análise espectral de sequências de DNA é mapear os dados simbólicos para um sinal numérico. Porém, o espectro dessas sequências é sensível ao mapeamento, portanto, idealmente, cada sequência de DNA deve ser mapeada para um sinal usando um

mapeamento específico, de modo que esse sinal capture o máximo possível de informações sobre a sequência. Em um esquema de mapeamento adaptativo, um espaço de busca com mapeamentos potenciais é delimitado e alguma condição pré-estabelecida deverá determinar qual o mapeamento que melhor evidencia a estrutura dos dados de uma sequência de DNA. Para implementar o método de mapeamento adaptativo, propomos o uso da abordagem de envoltória espectral.

A envoltória espectral representa a energia máxima de um sinal

$$\mathbf{x} = a\mathbf{x}_{\mathtt{A}} + c\mathbf{x}_{\mathtt{C}} + g\mathbf{x}_{\mathtt{G}} + t\mathbf{x}_{\mathtt{T}} = \mathbf{w}\left[\mathbf{x}_{\mathtt{A}} \;\; \mathbf{x}_{\mathtt{C}} \;\; \mathbf{x}_{\mathtt{G}} \;\; \mathbf{x}_{\mathtt{T}}\right]^T, \tag{5.1}$$

tal que $\mathbf{w} = [a \;\; c \;\; g \;\; t]$, $||\mathbf{w}|| = 1$ e sua energia espectral é dada por:

$$S[k] = \left|aX_{\mathtt{A}}[k] + cX_{\mathtt{C}}[k] + gX_{\mathtt{G}}[k] + tX_{\mathtt{T}}[k]\right|^2. \tag{5.2}$$

Para cada frequência no intervalo $k \in [0, N-1]$, existe um respectivo $\mathbf{w}$. Esses vetores são o espaço de busca para nosso método de mapeamento adaptativo baseado na envoltória espectral. Para cada $\mathbf{w}$, existe um mapeamento associado $\mathcal{M}$. Isto é, a imagem de um determinado mapeamento $\mathcal{M}$, ou seja, $a,\ c,\ g,\ t \in \mathbb{C}$, são os componentes do respectivo vetor $\mathbf{w}$. Portanto, no espaço de busca, existem até N mapeamentos potenciais e N sinais diferentes, que também podem diferir em sua composição espectral.

Por exemplo, considerando a sequência de codificação do gene AIM41 (geneID: 854425) do cromossomo XV de *Saccharomyces cerevisiae* com $N = 558$ bp. Por ser uma sequência codificante, esperamos a presença da propriedade TBP e, portanto, esperamos que seu espectro de energia revele um pico espectral discriminante na frequência $1/3$ rad/amostra. A envoltória espectral para esta sequência é mostrada na Figura 5.1(a) cuja região sombreada corresponde às frequências $1/3 \pm 0,02$ rad/amostra. É possível observar que, contradizendo o que seria esperado pela propriedade TBP, para a envoltória espectral o pico ocorre em $k = 0,22$ rad/amostra. No entanto, quando resolvemos a envoltória espectral na frequência $k_1 = 1/3$ rad/amostra, obtemos $\lambda_{k_1} = 2157,64$ e $\mathbf{w}_{k_1} = [0,43 \;\; -0,25+0,29j \;\; -0,13-0,69j \;\; -0,05+0,41j]$. Portanto, o mapeamento correspondente é dado por

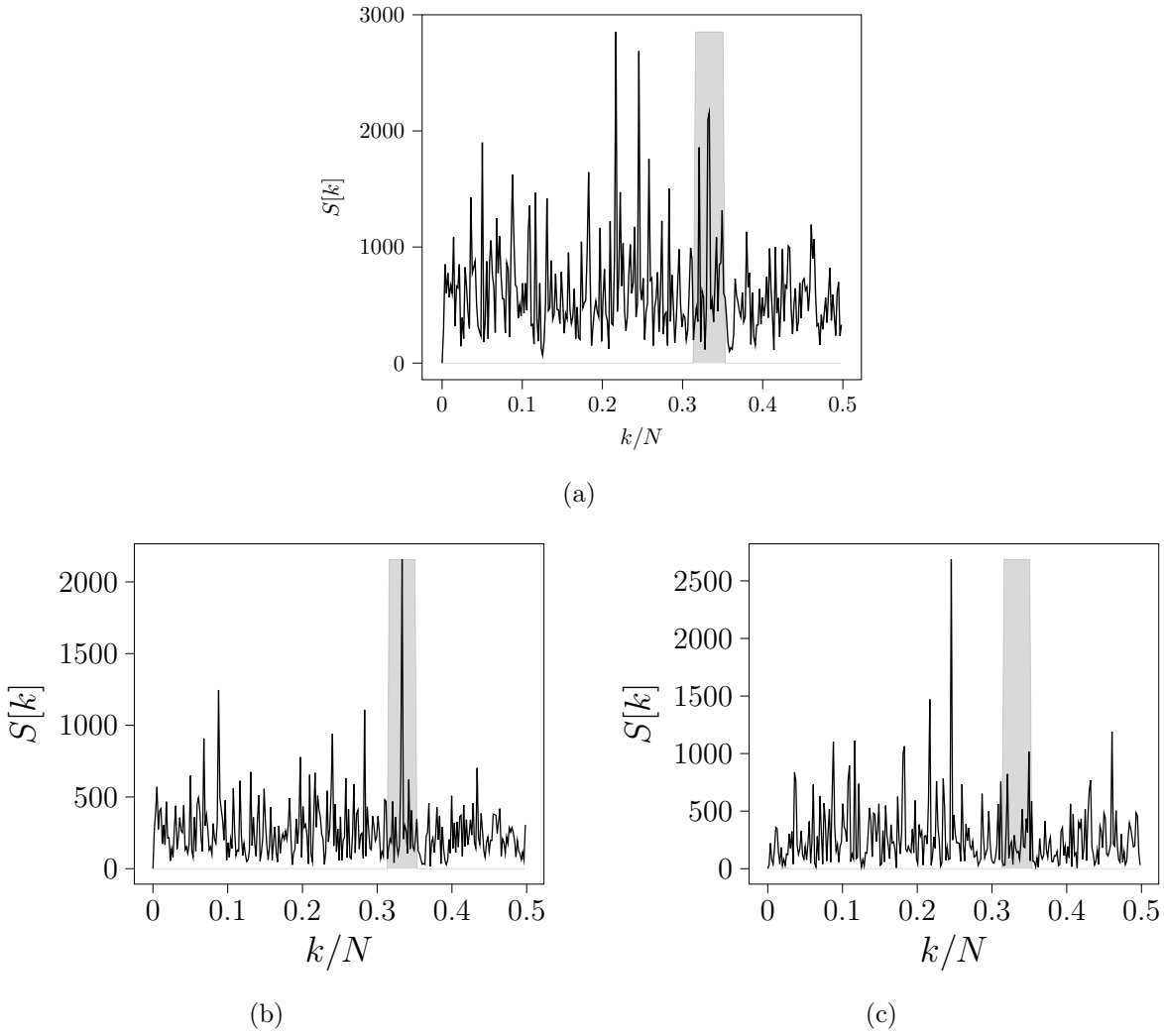
$$\begin{aligned} \mathcal{M}_1: \;\; & \mathtt{A} \mapsto 0,43,\ \mathtt{C} \mapsto -0,25+0,29j, \\ & \mathtt{G} \mapsto -0,13-0,69j,\ \mathtt{T} \mapsto -0,05+0,41j. \end{aligned} \tag{5.3}$$

O espectro de energia do sinal mapeado usando $\mathcal{M}_1$ é mostrado na Figura 5.1(b). Neste caso, como esperado pela propriedade TBP, o pico ocorre em $k = 0,33$ rad/amostra.

Porém, a propriedade TBP não é observada para todos os $\mathbf{w}$ no espaço de busca da envoltória espectral. É o que podemos observar para o mapeamento correspondente da envoltória espectral na frequência $k_2 = 0,24$ rad/amostra. Nesse caso, a energia da envoltória espectral é $\lambda_{k_2} = 2687,59$ e o mapeamento correspondente é dado por

$$\begin{aligned} \mathcal{M}_2: \;\; & \mathtt{A} \mapsto 0,41,\ \mathtt{C} \mapsto 0,48+0,23j, \\ & \mathtt{G} \mapsto -0,29-0,07j,\ \mathtt{T} \mapsto -0,06+0,29j. \end{aligned} \tag{5.4}$$

Figura 5.1 – Espetro de energia do gene AIM41 (geneID: 854390) do cromossomo XV de *Saccharomyces cerevisiae* com $N = 558$ bp. (a) Envoltória espectral. (b) Espectro de energia do sinal resultante do mapeamento usando $\mathcal{M}_1$ como na Equação (5.3). (c) Espectro de energia do sinal resultante do mapeamento usando $\mathcal{M}_2$ como na Equação (5.4).



Fonte: Elaborada pela autora.

O espectro de energia do sinal mapeado usando $\mathcal{M}_2$ é mostrado na Figura 5.1(c). Mais uma vez é possível observar que o espectro de energia das sequências de DNA podem ser ligeiramente diferentes quando alteramos o mapeamento.

Para selecionar um único mapeamento para uma sequência de DNA, devemos escolhê-lo entre os N mapeamentos potenciais resultantes da envoltória espectral. Por esta razão, uma restrição deve ser imposta. O primeiro algoritmo usa como restrição a maximização da SNR do espectro de densidade de energia. Consequentemente, a partir de agora, vamos denominá-lo de SNR-SE, em que SE é a forma abreviada para envoltória espectral. A SNR é a razão entre a potência do sinal e a potência do ruído que são estimadas a partir do espectro de energia do sinal como segue. A potência do sinal é estimada como a energia da componente espectral mais alta; a potência do ruído ou o

Algoritmo 3 SNR-SE

Entrada: Sequência de DNA s
Saída: Espectro unilateral $S[k]$

$snr_{ref} \leftarrow -\infty$
para cada k em $[0, \lfloor N/2 \rfloor]$ **faça**
  $\mathbf{w} \leftarrow$ SpectralEnvelope (s, k)
  $\mathcal{M}_{ref} \leftarrow$ mapa cuja imagem são as componentes de $\mathbf{w}$
  Calcula $S[k]$ usando a Equação (5.2) para o mapeamento $\mathcal{M}_{ref}$
  $snr \leftarrow \mathrm{SNR}(S[k])$
  se $snr_{ref} < snr$ **então**
    $snr_{ref} \leftarrow snr$
    $\mathcal{M} \leftarrow \mathcal{M}_{ref}$
  fim se
fim para
Calcula $S[k]$ usando a Equação (5.2) para o mapeamento $\mathcal{M}$
retorna $S[k]$

Algoritmo 4 TBP-SE

Entrada: Sequência de DNA s
Saída: Espectro unilateral $S[k]$

$k \leftarrow \lfloor N/3 \rfloor$
$\mathbf{w} \leftarrow$ SpectralEnvelope (s, k)
$\mathcal{M} \leftarrow$ mapa cuja imagem são as componentes de $\mathbf{w}$
Calcule $S[k]$ usando a Equação (5.2) para o mapeamento $\mathcal{M}$
retorna $S[k]$

ruído de fundo é a energia total, excluindo a potência do sinal e o valor médio [6].

Neste algoritmo, os mapeamentos do espaço de busca serão aqueles que resolvem a envoltória espectral para cada frequência k no intervalo $[0, \lfloor N/2 \rfloor]$. O espaço de busca é reduzido, pois restringe-se a análise do espectro de energia unilateral, que deve conter toda a informação espectral sobre o sinal. Portanto, para cada mapeamento potencial, estima-se o espectro de energia e sua SNR. Por fim, escolhemos o mapeamento cujo respectivo sinal possui o espectro de energia que maximiza a SNR. O pseudocódigo deste método é apresentado no Algoritmo 3. O pseudocódigo do algoritmo SpectralEnvelope (s, k) foi apresentado no Algoritmo 1.

O segundo algoritmo é um caso especial do primeiro. Nesse método, propõe-se explorar o conhecimento prévio da propriedade TBP. Por esta razão, a partir de agora o denominaremos de TBP-SE. Nesse caso, assume-se que todas as sequências codificantes possuem a propriedade TBP, de modo que um pico espectral discriminante na frequência $1/3$ rad/amostra é sempre observado, enquanto que, em sequências não codificantes, esse pico está ausente. Portanto, dentre os mapeamentos potenciais da envoltória espectral, este algoritmo escolhe aquele que resolve o problema da envoltória espectral na frequência $k = \lfloor N/3 \rfloor$ rad/amostra. O pseudocódigo deste método é apresentado no Algoritmo 4.

5.2 Características dos Algoritmos

Mapeamento Complexo

Como os algoritmos propostos buscam mapeamentos complexos, o espectro de densidade de energia pode ser assimétrico no eixo de frequência. Por esta razão, é extremamente importante levar em consideração tanto o conteúdo das frequências positivas quanto o das negativas para determinar o espectro unilateral. Considerando, por exemplo, a sequência periódica (com periodicidades de 3 bp e 6 bp) da qual o primeiro período é dado por:

$$s = \texttt{CACCCG}\cdots . \tag{5.5}$$

De fato, pode-se verificar as periodicidades dessa sequência considerando o seguinte sinal senoidal,

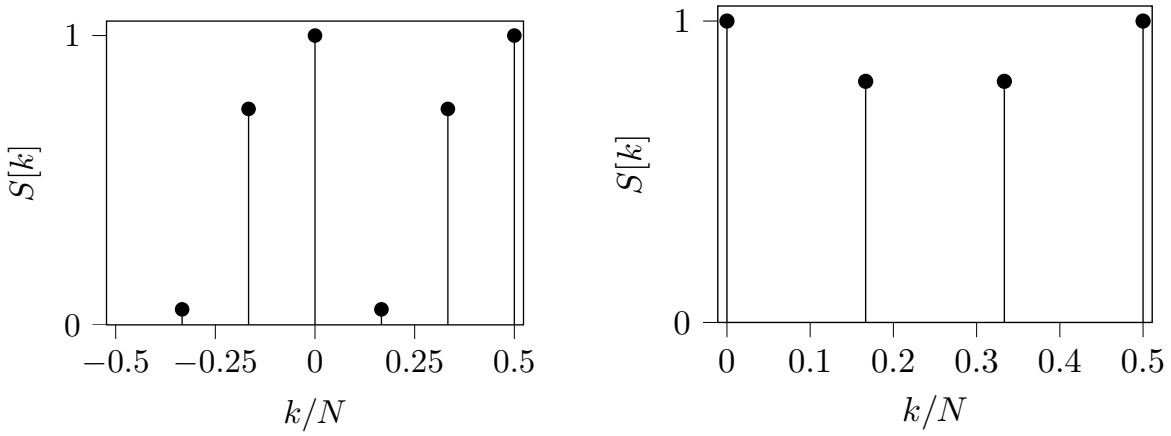
$$x[n] = \sin\left(\frac{2\pi}{3}n\right) + \sin\left(\frac{2\pi}{6}n\right), \tag{5.6}$$

em que $x[n]$ assume apenas três valores: $-\sqrt{3}$, 0 e $\sqrt{3}$. Ao definir $\mathcal{M} : \texttt{A} \mapsto \sqrt{3}$, $\texttt{C} \mapsto 0$, $\texttt{G} \mapsto -\sqrt{3}$, $\texttt{T} \mapsto t$ em que t pode assumir qualquer valor, o sinal resultante do mapeamento de s com $\mathcal{M}$ é dado na Equação (5.6).

Portanto, espera-se que o espectro de de energia de s tenha picos que ocorram nas frequências $k_1 = 1/6$ rad/amostra e $k_2 = 1/3$ rad/amostra. Na Figura 5.2(a) é possível ver o espectro de energia de s quando o mapeamento QPSK foi realizado. Esse espectro é assimétrico com relação ao eixo das frequências e tem outros dois picos em $k = 0$ rad/amostra e $k = 1/2$ rad/amostra. Uma vez que esses picos são tradicionalmente desconsiderados na análise espectral, analisa-se apenas as demais componentes de frequência existentes. Observa-se que, ao analisar apenas o conteúdo das frequências positivas, o pico em k_1 é menor que o pico em k_2; e se analisarmos apenas o conteúdo das frequências negativas, o pico em $-k_1$ é maior que o pico em $-k_2$; no entanto, era esperado que ambos os picos tivessem o mesmo conteúdo espectral. Portanto, para uma análise confiável, o espectro unilateral deve ser calculado adicionando o conteúdo espectral das frequências negativas ao conteúdo espectral nas frequências positivas. Finalmente, o espectro unilateral de s, usando o mapeamento QPSK, é mostrado na Figura 5.2(b). Nesse caso, ambos os picos têm o mesmo conteúdo.

Para esta sequência específica, nossos algoritmos também encontraram um mapeamento complexo. O algoritmo TBP-SE retorna $\mathcal{M} : \texttt{A} \mapsto 0,58$, $\texttt{C} \mapsto -0,29 - 0,5j$, $\texttt{G} \mapsto -0,29 + 0,5j$ e $\texttt{T} \mapsto 0$; e o SNR-SE retorna $\mathcal{M} : \texttt{A} \mapsto 0,50$, $\texttt{C} \mapsto -0,71 + 0,2j$, $\texttt{G} \mapsto -0,45 - 0,004j$ e $\texttt{T} \mapsto 0$. Ambos produzem o mesmo espectro de densidade de energia e estão em conformidade com o espectro mostrado na Figura 5.2(b). Uma análise similar pode ser feita para qualquer outra sequência de DNA usando mapeamentos complexos.

Figura 5.2 – Espectro de energia usando o mapeamento QPSK para a sequência s definida na Equação (5.5): (a) Espectro bilateral: existem picos nas frequências $k_1 = 1/6$ rad/amostra e $k_2 = 1/3$ rad/amostra, mas com conteúdos diferentes. (b) Espectro unilateral: existem picos nas frequências $k_1 = 1/6$ rad/amostra e $k_2 = 1/3$ rad/amostra com o mesmo conteúdo.



(a) (b)

Fonte: Elaborada pela autora.

Mapeamento Adaptativo

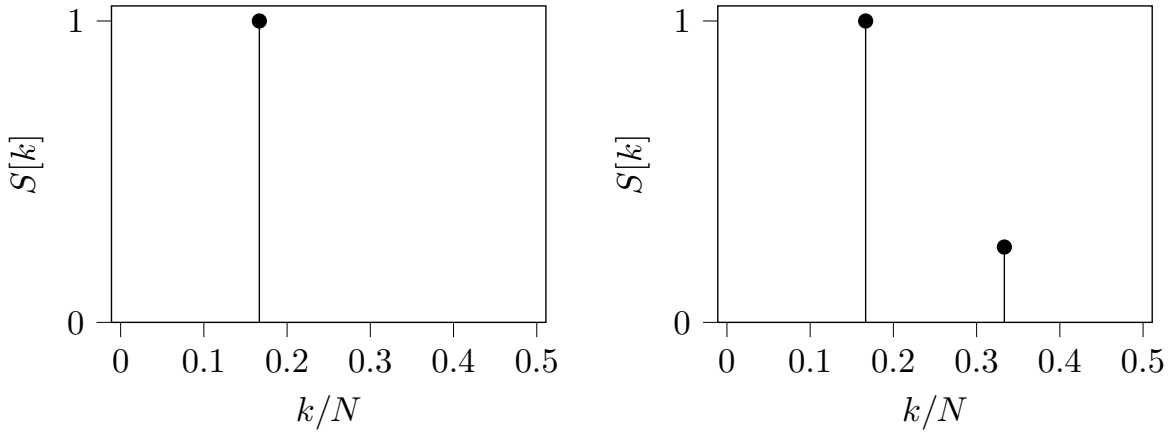
As sequências simbólicas podem ter uma estrutura estatística que fornece informações importantes sobre elas. Portanto, um mapeamento do domínio simbólico para o numérico não deve adicionar informação à sequência simbólica além do que já é inerente a ela. Por exemplo, um mapeamento arbitrário seria atribuir as bases do DNA ordenadas alfabeticamente a uma sequência crescente de números inteiros, como segue, $\mathcal{M} : \mathtt{A} \mapsto 1$, $\mathtt{C} \mapsto 2$, $\mathtt{G} \mapsto 3$ e $\mathtt{T} \mapsto 4$. Contudo, este $\mathcal{M}$ sugere que uma base é de alguma forma maior que outra e, esta é uma propriedade que este conjunto simbólico não possui [56].

Outro exemplo é a sequência periódica cujo primeiro período é mostrado a seguir,

$$s = \mathtt{ACGTGC}\cdots . \tag{5.7}$$

Nesse caso, o espectro de energia de s muda significativamente dependendo do mapeamento escolhido. Por exemplo, quando $\mathcal{M}_1 : \mathtt{A} \mapsto 1$, $\mathtt{C} \mapsto 0,5$, $\mathtt{G} \mapsto -0,5$, $\mathtt{T} \mapsto -1$, então $S[k]$ tem apenas um pico em $k = 1/6$ rad/amostra (Figura 5.3(a)). No entando, se $\mathcal{M}_2 : \mathtt{A} \mapsto 1,5$, $\mathtt{C} \mapsto 0,25$, $\mathtt{G} \mapsto -0,75$, $\mathtt{T} \mapsto -0,5$, então, $S[k]$ tem um pico em $k = 1/6$ rad/amostra e um pico em $k = 1/3$ rad/amostra (Figura 5.3(b)). É possível observar que dependendo do mapeamento, podemos detectar ou não a periodicidade na frequência $1/3$ rad/sample. Nesse caso, o espectro de energia dos sinais resultantes de diferentes mapeamentos nem sempre revelaram toda a informação espectral contida na sequência simbólica. Essa é mais uma razão para enfatizar a importância de ter uma flexibilidade no mapeamento, uma vez que não parece haver um mapeamento único e adequado para análise espectral de todas as sequências de DNA.

Figura 5.3 – Espectro de energia unilateral da sequência s definida na Equação (5.7) quando: (a) $\mathcal{M}_1$: `A` $\mapsto 1$, `C` $\mapsto 0.5$, `G` $\mapsto -0.5$ e `T` $\mapsto -1$ é usado; e (b) $\mathcal{M}_2$: `A` $\mapsto 1.5$, `C` $\mapsto 0.25$, `G` $\mapsto -0.75$ e `T` $\mapsto -0.5$ é usado.



(a) Mapeamento $\mathcal{M}_1$. (b) Mapeamento $\mathcal{M}_2$.

Fonte: Elaborada pela autora.

Por outro lado, nos algoritmos propostos na seção anterior, o mapeamento atua como um parâmetro e é escolhido unicamente para cada sequência de acordo com sus propriedades espectrais. Este também é o caso do espectro MEM [8], no qual, o critério de minimização da entropia espectral é usado. No entanto, a entropia espectral é invariante sob a permutação das estimativas do espectro de energia na faixa de frequência, ignorando assim a estrutura de ordem parcial intrínseca de um sinal [86]. Nesse caso, sinais muito diferentes no domínio do tempo produzem a mesma entropia espectral. Portanto, este critério de otimização pode resultar na perda de informação sobre o sinal.

Explorando a Propriedade TBP

Especialmente em aplicações que lidamos com sequências simbólicas e há conhecimento prévio sobre suas características espectrais, podemos utilizar tais informações para melhorar a análise. Por exemplo, sabemos que a propriedade TBP está presente em regiões exônicas e ausente em regiões intrônicas, portanto devemos verificar se é possível maximizar este conteúdo de frequência para melhorar a discriminação das regiões do DNA. Isso é exatamente o que é proposto no algoritmo TBP-SE.

Complexidade Computacional

A complexidade computacional dos algoritmos de análise espectral discutidos neste capítulo será avaliada no sentido da notação *big O*. Essa notação é particularmente útil para estudar o comportamento de um algoritmo sob as condições extremas, em que estamos frequentemente satisfeitos com um limite superior nos recursos consumidos pelo

algoritmo [99].

Observe que o cálculo da DFT das quatro funções indicadoras binárias é a operação comum para todos os algoritmos. Esta operação tem complexidade $O(N \log N)$, em que N é o comprimento da sequência. Em alguns métodos, este é o termo com a ordem mais alta, e por isso dizemos que Voss [53], EIIP [54], QPSK [5], SNR-SE e TBP- SE são $O(N \log N)$. Por outro lado, as operações adicionais necessárias no MEM Spectrum [8] têm ordem quadrática, então o espectro MEM é $O(N^2)$.

5.3 Método de Avaliação Estatística

A classificação de sequências de DNA codificantes a partir da análise espectral é realizada da seguinte maneira. Verifica-se em qual frequência ocorre o maior pico espectral. Se este pico ocorrer entre as frequências $1/3 \pm 0,02$ rad/amostra, classificamos tal sequência como uma sequência codificadora de proteína. Caso contrário, tal sequência é classificada como não codificadora.

À priori, o resultado desse teste pode ser positivo (classificando a sequência de DNA como uma sequência codificante) ou negativo (classificando a sequência de DNA como uma sequência não codificante). Porém, esses resultados podem ou não corresponder a condição real. Portanto, têm-se que esses testes podem ter os seguintes resultados:

- Verdadeiro positivo (TP): sequências codificantes identificadas corretamente como sequências codificantes;
- Falso positivo (FP): sequências não codificantes que são classificadas erroneamente como sequências codificantes;
- Verdadeiro negativo (TN): sequências não codificantes que são corretamente classificadas como sequências não codificantes;
- Falso negativo (FN): sequências codificantes que são classificadas erroneamente como sequências não codificantes.

Para comparar a eficácia de cada mapeamento para a identificação de sequência codificantes de DNA, avaliamos três medidas: acurácia, sensibilidade e especificidade. A acurácia avalia a taxa de classificação correta global, refletindo a capacidade de prever corretamente em relação ao total de amostras, ou seja,

$$\textit{acurácia} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{5.8}$$

A sensibilidade ou taxa de verdadeiro positivo (TPR, do inglês, *True Positive Rate*) avalia a capacidade do classificador de predizer uma sequência codificadora de proteína corretamente, ou seja,

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{5.9}$$

A especificidade ou taxa de verdadeiro negativo (TNR, do inglês, *True Negative Rate*) avalia a capacidade do classificador de predizer uma sequência não codificadora corretamente, ou seja,

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{5.10}$$

De forma geral, se o teste indicar que a sensibilidade é alta, quer dizer que o método também tem alta probabilidade de classificar uma sequência de DNA que realmente é uma sequência codificante como uma sequência codificante; assim, o evento de classificar uma sequência codificante erroneamente como sequência não codificante é raro (ou seja, baixo FN). A mesma linha de raciocínio segue para a interpretação da especificidade. Quando a especificidade é alta, quer dizer que o método também tem alta probabilidade de classificar uma sequência de DNA que realmente não é uma sequência codificantes como uma sequência não codificante; assim, o evento de classificar uma sequência não codificante como sequência codificante é raro (ou seja, baixo FP). O melhor método de previsão possível produz o seguinte resultado: sensibilidade de 100% (não existem falsos negativos) e especificidade de 100% (não existem falsos positivos).

5.4 Resultados

Dados Experimentais

Para uma análise detalhada dos espectros de energia para diferentes mapeamentos, usamos os cromossomos XIV, XV e XVI de *Saccharomyces cerevisiae* (números de acesso NC_001146.8, NC_001147.6 e NC_001148.4, respectivamente). Cada cromossomo tem 398, 546 e 474 sequências codificantes, respectivamente. Para a sequências codificantes cuja orientação era complementar, realizamos a operação de complemento reverso para iniciar cada sequência no códon `ATG`. Os dados são divididos em dois conjuntos de dados: o primeiro possui apenas sequências codificantes (o conjunto de dados de sequência codificantes) e o segundo possui apenas sequências de regiões intergênicas (o conjunto de dados de sequência não codificante). Em ambos os casos, descartamos sequências cujo comprimento é menor que 200 bp. Por fim, existem 1388 sequências codificantes e 1188 sequências não codificantes em nosso conjunto de dados. A análise estatística descrita na próxima seção irá considerar esse conjunto de dados.

Além disso, utilizamos a porção do gene F56F11 do cromossomo III de *Caenorhabditis elegans* que transcreve a proteína F56F11.4, isoforma a. O F56F11.4a é usado como um problema de referência para diferentes técnicas de detecção de éxon [5, 8, 100] e possui 7990 bp começando na posição 7021 do gene F56F11. Além disso, o F56F11.4a tem cinco éxons distintos bem conhecidos cujas localizações relativas à posição do nucleotídeo 7021 variam de 928 a 1039, 2528 a 2857, 4114 a 4377, 5465 a 5644 e 7255 a 7605.

Análise Estatística

Conforme visto até então, o espectro de energia das sequências de DNA pode ser ligeiramente diferente quando comparamos diferentes representações numéricas para a análise espectral. Em geral, esses espectros não representam versões aproximadas entre si. Para comparação, o espectro de energia de todas as sequências do nosso conjunto de dados foi avaliado usando os dois algoritmos propostos neste capítulo: SNR-SE e TBP-SE, bem como com outros quatro métodos já consolidados na literatura: Voss [53], EIIP [54], QPSK [5] e MEM spectrum [8].

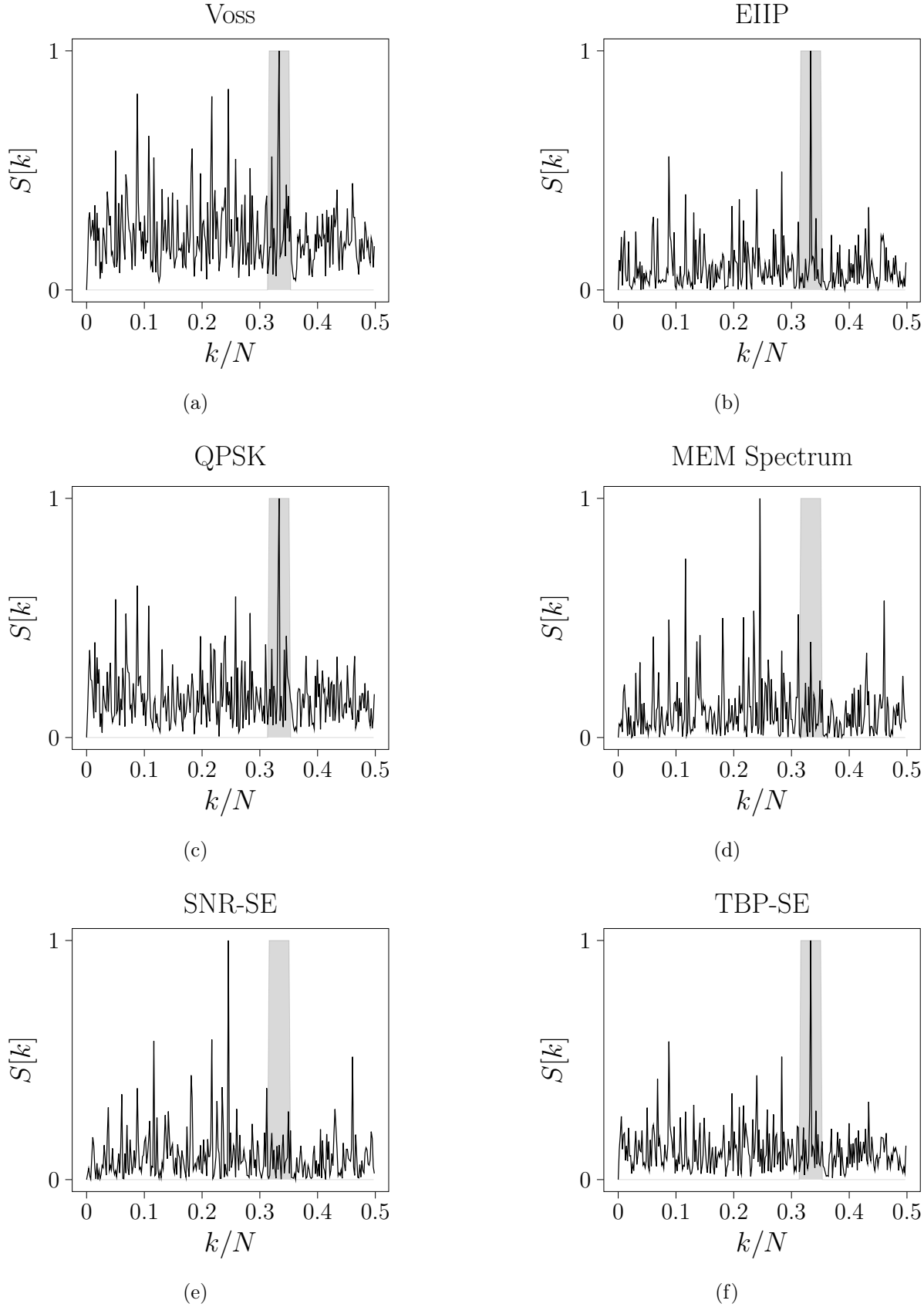
Considerando o caso específico dos genes AIM41 e MPR35 cujos espectros de energia são mostrados nas Figuras 5.4 e 5.5, respectivamente, cujas regiões sombreadas correspondem às frequências $1/3 \pm 0,02$ rad/amostra. É possível observar que, conforme esperado, nem todos os métodos detectam a propriedade TBP para esses genes. Existem duas possíveis razões para isso. Primeiro, o mapeamento escolhido pode ocultar informações espectrais sobre a sequência. Para o gene AIM41, por exemplo, o espectro de densidade de energia, conforme definido por Voss ou usando mapeamentos EIIP, QPSK e TBP-SE, tem o maior pico na frequência $0,33$ rad/amostra. No entanto, o ruído de fundo aumenta significativamente quando o Voss é avaliado. Além disso, esta frequência discriminatória é perdida quando o espectro MEM e SNR-SE são avaliados.

A segunda razão é que, embora a propriedade TBP em sequências de codificação seja um discriminador de frequência clássico no contexto biológico, algumas sequências de codificação não são distinguidas pela mesma. Esse é o caso do gene MPR35. Para todos os métodos, o espectro de densidade de energia tem o maior pico na frequência $0,09$ rad/amostra. Porém, mesmo considerando a existência desses casos, em geral, o espectro obtido pelos métodos propostos por nós produz não apenas melhorias na classificação da sequência de codificação, como também na redução do ruído do espectro.

Embora existam limitações intrínsecas na análise espectral de uma determinada sequência de DNA, alguns métodos podem discriminar melhor a propriedade TBP para sequências de codificação do que outros. A Tabela 5.1 compara os métodos mencionados em relação à acurácia, sensibilidade e especificidade para o conjunto de sequências dos cromossomos XIV, XV e XVI de *Saccharomyces cerevisiae* descrito anteriormente. O *trade-off* entre sensibilidade e especificidade também se aplica nesse caso, de modo que, aumentando a sensibilidade, diminui-se a especificidade e vice-versa.

A Tabela 5.1 revela que o método proposto, TBP-SE, teve a maior acurácia e sensibilidade entre todos. Isso é especialmente importante nesta aplicação, pois reduzimos a probabilidade de que uma sequência de codificação não seja identificada. Em outras palavras, é mais provável que as sequências de codificação sejam identificadas corretamente como sequências de codificação usando TBP-SE. Além disso, a especificidade teve um nível expressivo, e TBP-SE teve os níveis mais uniformes para as três métricas: acurácia, sensibilidade e especificidade.

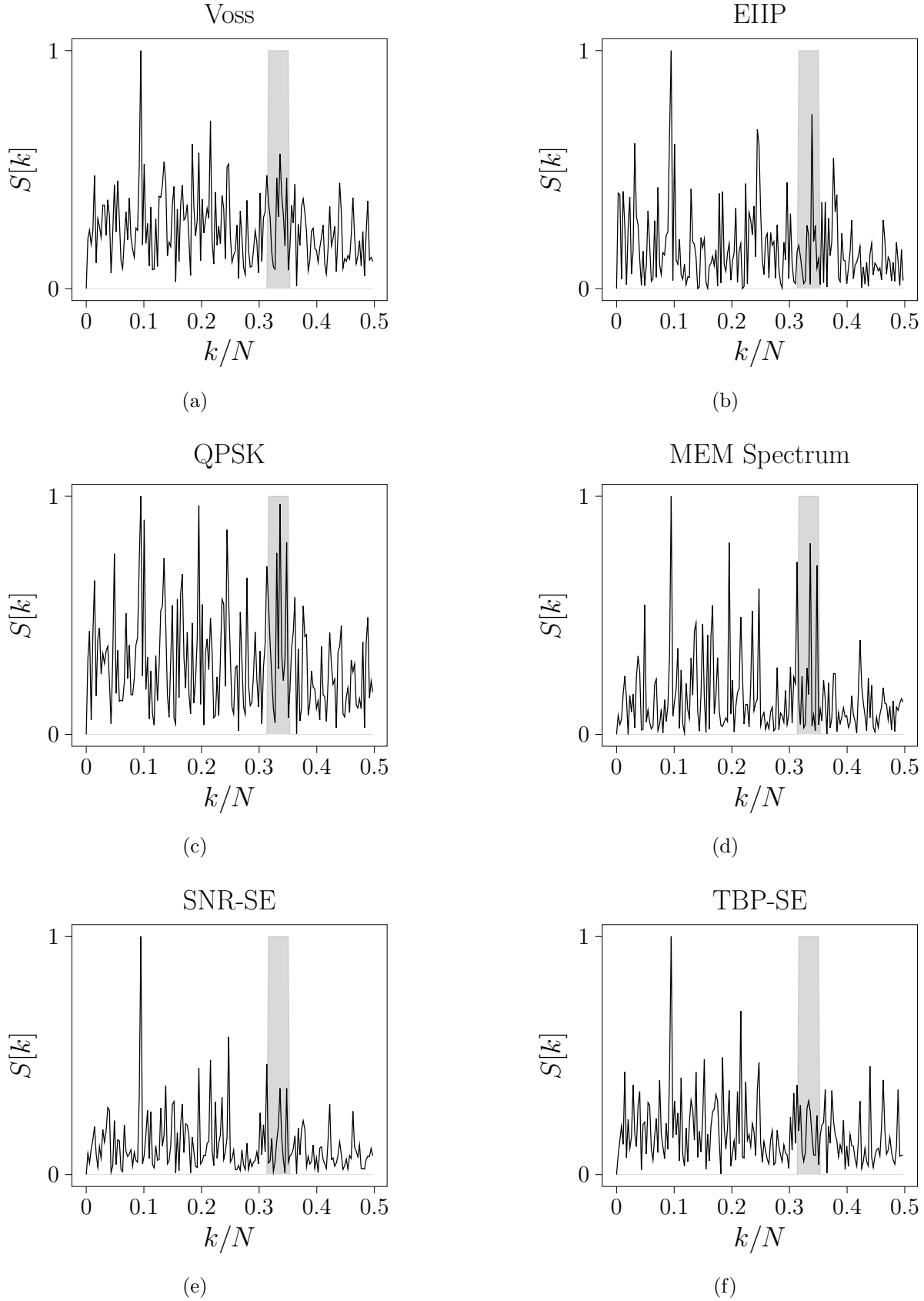
Figura 5.4 – Espectro de energia normalizado para a CDS do gene AIM41 (geneID: 854390) do cromossomo XV de *S. cerevisiae* com $N = 558$ bp. (a) Voss. (b) EIIP. (c) QPSK. (d) MEM Spectrum. (e) SNR-SE. (f) TBP-SE.



(a) (b) (c) (d) (e) (f)

Fonte: Elaborada pela autora.

Figura 5.5 – Espectro de energia normalizado para a CDS do gene MRP35 (geneID: 855601) do cromossomo XIV de *S. cerevisiae* com $N = 348$ bp. (a) Voss. (b) EIIP. (c) QPSK. (d) MEM Spectrum. (e) SNR-SE. (f) TBP-SE.



(a) (b) (c) (d) (e) (f)

Fonte: Elaborada pela autora.

Tabela 5.1 – Taxa de discriminação entre sequências de DNA codificadoras e não codificadoras por meio da análise espectral.

Método	**Acurácia (%)**	**TPR (%)**	**TNR (%)**
Voss [53]	88,00	80,61	96,63
EIIP [54]	85,79	81,77	90,48
QPSK [5]	86,25	78,96	94,78
MEM [8]	74,84	59,43	92,84
SNR-SE	86,64	79,97	94,44
TBP-SE	90,41	89,26	91,74

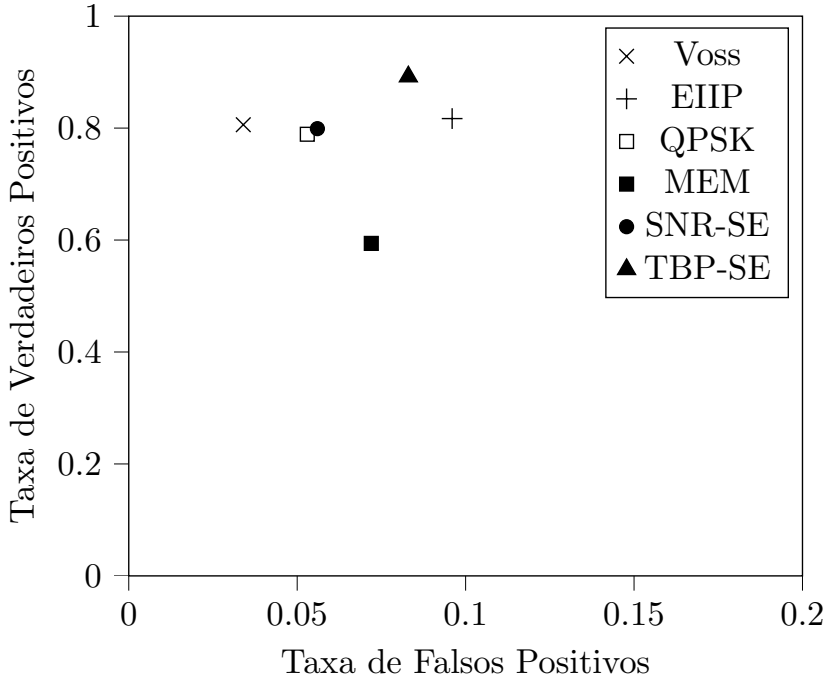
Por outro lado, ao comparar os métodos de mapeamento adaptativo, o espectro utilizando-se MEM não apresenta um bom desempenho. Esse mapeamento apresentou os níveis mais baixos de acurácia e sensibilidade. Uma possível razão para isso é que o espaço de busca desse método é limitado pela entropia espectral; no entanto, a entropia espectral ignora a estrutura intrínseca da ordem parcial, como indicado por [86]. Além disso, tal método possui a maior complexidade computacional e não é viável quando comparado aos outros métodos de análise espectral discutidos neste artigo.

Embora os métodos Voss, EIIP, QPSK e SNR-SE pareçam ter um desempenho semelhante, as diferenças podem ser notadas graficamente por meio da curva Característica de Operação do Receptor (ROC, do inglês: *Receiver Operating Characteristic*). A curva ROC de um classificador ideal deve indicar uma alta TPR e uma baixa taxa de falsos positivos (FPR, do inglês: *False Positive Rate*), em que $\text{FPR} = 1 - \text{TNR}$. Como usamos uma classificação binária, as estatísticas do método produzem um único ponto no espaço ROC.

A curva ROC para o nosso experimento revela informações importantes sobre o desempenho na identificação de sequências codificantes observando o espectro de energia dos sinais resultantes da representação numérica dessas sequências. A curva ROC para essa classificação é apresentada na Figura 5.6. O desempenho para QPSK e SNR-SE é semelhante (a diferença em TPR é 0,01 e em FPR é 0,003). Além disso, Voss, EIIP, QPSK e SNR-SE têm aproximadamente o mesmo TPR, porém, Voss é preferido porque tem o FPR mais baixo. Observa-se, ainda, que o Voss e o TBP-SE estão situados na fronteira de Pareto, e, portanto, dominam os demais métodos.

Ao comparar Voss e TBP-SE observou-se que aproximadamente 97,31% de sequências de codificação que foram classificadas erroneamente como sequências não codificantes usando TBP-SE também foram classificadas erroneamente usando Voss. Este fenômeno ocorre, por exemplo, no gene MRP35 (Figura 5.5). Porém, ainda assim, o ruído do espectro de DNA é reduzido usando TBP-SE. Portanto, TBP-SE pode ser preferível a Voss, já que a métrica TPR é especialmente importante nesta aplicação.

Figura 5.6 – Curva ROC para a classificação por meio da análise espectral de sequências de DNA.



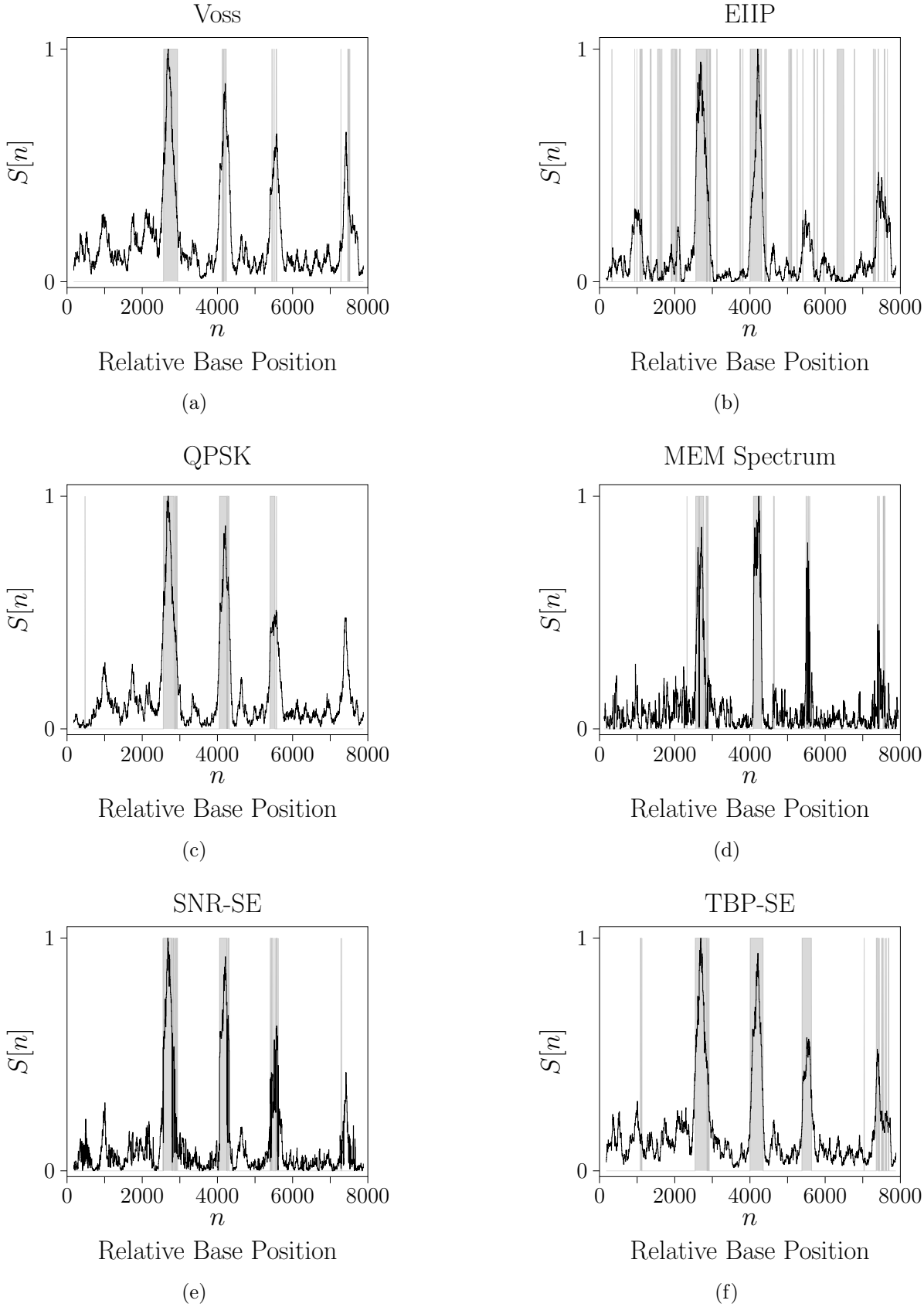
Fonte: Elaborada pela autora.

Estudo de caso: Gene F56F11.4a

Conforme já exposto anteriormente, o gene F56F11.4a tem cinco éxons distintos bem conhecidos cujas localizações estão entre 928 e 1039, 2528 e 2857, 4114 e 4377, 5465 e 5644, e 7255 e 7605. O primeiro éxon é o mais curto (112 bases) e geralmente é o mais difícil de detectar. Neste cenário, as regiões de codificação são identificadas analisando o espectro de energia no domínio tempo-frequência [4, 100, 14]. Ou seja, o espectro de densidade de energia na frequência $1/3$ rad/amostra é avaliado em uma janela de W amostras, então a janela é deslizada ao passo de uma ou mais amostras e a densidade de energia é recalculada em um processo que analisa toda a sequência de DNA. Um critério importante para esta análise é definir o comprimento da janela W. Para este gene, Tiwari et al. [4] sugere usar $W = 351$. Portanto, uma janela retangular de comprimento 351 e tamanho de passo 5 foi usada.

Os resultados são apresentados na Figura 5.7, em que o eixo horizontal são as posições relativas da base e o eixo vertical é o espectro de energia normalizado pelo seu valor máximo. Duas interpretações são possíveis nesse cenário. Primeiro, os picos no espectro devem corresponder às regiões onde a propriedade TBP está presente. Essas regiões podem ser avaliadas usando um limiar, ou seja, as regiões codificantes são identificadas colocando um limiar no espectro, de modo que as regiões com energia acima desse limiar são consideradas éxons; e as regiões com energia abaixo desse limiar são consideradas íntrons. Neste caso, em geral, os métodos detectam quatro dos cinco éxons e o primeiro

Figura 5.7 – Espectro de energia janelado do gene F56F11.4a usando janela de tamanho $W = 351$ para os seguintes métodos: (a) Voss; (b) EIIP; (c) QPSK; (d) MEM spectrum; (e) SNR-SE; (f) TBP-SE.



Fonte: Elaborada pela autora.

é, geralmente, esquecido. Especificamente para o método EIIP, a energia do quarto éxon é significativamente reduzida e se mistura com regiões intrônicas. Os outros métodos têm desempenho semelhante.

No entanto, a segunda interpretação desses resultados extrai mais informações sobre a localização de regões exônicas neste gene. Neste caso, observa-se para cada janela se a propriedade TBP está presente. Na Figura 5.7, a presença da propriedade TBP é indicada pelas regiões sombreadas. É importante destacar que o sinal resultante do mapeamento TBE-SE foi o único a identificar a presença de todos os cinco éxons. O EIIP de fato mostrou ter mais instabilidade na previsão de regiões não codificantes. Voss, QPSK, MEM e SNR-SE tiveram desempenhos semelhantes, porém, o MEM aumenta o ruído de fundo do espectro de energia. Embora o QPSK pareça detectar um éxon adicional no início da sequência, essa área sombreada está localizada longe da região do primeiro éxon verdadeiro. Além disso, não existem áreas sombreadas na região do último éxon. Todos esses resultados eram esperados com base na análise prévia da curva ROC.

5.5 Considerações

As sequências de DNA podem ser interpretadas como sequências simbólicas e, portanto, idealmente, sua representação numérica não deve impor características adicionais ao sinal mapeado. Como visto anteriormente, o espectro desses sinais é sensível ao mapeamento. Isto é, considerando os sinais resultantes de mapeamentos distintos, os respectivos espectros de energia de uma dada sequência de DNA também são distintos, e não representam versões aproximadas um do outro. Além disso, a hipótese de considerar um mapeamento fixo para todas as sequências de DNA não garante que as características intrínsecas de cada sequência estará sendo devidamente extraída.

Por esta razão, idealmente, cada sequência de DNA deve ser mapeada para um sinal usando um mapeamento específico, de modo que esse sinal capture o máximo possível de informações sobre essa sequência. Portanto, neste capítulo, propomos dois algoritmos para definir regras de mapeamento para sequências de DNA usando a abordagem de envoltória espectral: SNR-SE e TBP-SE. Embora a envoltória espectral para sequências de DNA já tenha sido utilizado na literatura como um método para definir um novo espectro [56, 57], ele nunca foi usado para encontrar mapeamentos adaptativos para sequências de DNA.

Os algoritmos propostos são, portanto, novos métodos para encontrar mapeamentos complexos adaptativos para sequências de DNA e, assim, melhorar a análise espectral de tais sequências simbólicas. As observações sobre os algoritmos propostos são resumidas a seguir. A abordagem de envoltória espectral é usada para encontrar mapeamentos adaptativos e, assim, converter sequências de DNA em sinais de tempo discreto. Um mapeamento é escolhido exclusivamente para cada sequência observando condições de SNR e a propriedade TBP. O mapeamento foi definido sobre o corpo dos complexos.

Ambos os algoritmos possuem complexidade loglinear, ou seja, são $O(N \log N)$ em que N é o comprimento da sequência. A eficiência computacional é essencial quando longas sequências de DNA e bancos de dados de grande porte precisam ser processados.

Para investigar como nossos algoritmos melhoram a análise espectral do DNA para classificação de sequência de codificação de DNA, também verificamos a presença ou ausência da propriedade TBP no espectro de DNA para os seguintes métodos: Voss [53], EIIP [54] , QPSK [5], MEM [8]. Nesse cenário, o método proposto, TBP-SE, teve a maior acurácia e sensibilidade entre todos. Além disso, as abordagens TBP-SE e Voss apresentaram melhor desempenho para implementar esta classificação. No entanto, o TBP-SE se destaca, pois possui a maior sensibilidade, que é mais importante nesta aplicação, pois, assim, reduzimos a probabilidade de ter uma sequência de codificação que não será identificada. Também analisamos o desempenho dos métodos de identificação de regiões exônicas no gene F56F11.4a. Nesse caso, só foi possível identificar a presença de todos os cinco éxons do gene ao usar o mapeamento TBE-SE.

Parte IV

Conclusão

Capítulo 6

Considerações Finais

O processamento de sinais genômicos é uma grande área que compreende diversas aplicações. Nesse documento de Tese foram apresentados ao que concerne duas subáreas: teoria dos códigos (buscando por estruturas matemáticas adjacente às sequências de DNA) e análise espectral para discriminação de sequências codificantes e não codificantes de proteínas. Todo o embasamento teórico para o desenvolvimento da pesquisa em ambas as áreas de aplicações foi documentado nos Capítulos 2 e 3.

Especificamente, no Capítulo 2 foram apresentados os fundamentos teóricos de processamento da informação de uma forma geral. Já no Capítulo 3, esses fundamentos foram revistos e direcionados para o contexto da extração da informação genômica. Optou-se, portanto, a apresentar exemplos com sequências de DNA reais à medida que eram apresentadas técnicas de processamento genômico já existentes na literatura.

O Capítulo 4 foi dedicado à apresentação dos resultados referente ao estudo e análise de estruturas algébricas subjacentes às sequencias de DNA. Nesse contexto, foram propostas melhorias ao algoritmo *DNA Sequence Genarion* que resultaram em um novo algoritmo. A principal contribuição do novo algoritmo é a substituição do decodificador usando força bruta por um decodificador de síndrome. Constatou-se, portanto, a redução do número de operações realizadas no processo de busca de códigos BCH que identificam uma sequência de DNA. Diz-se que uma sequência de DNA é identificada por um código BCH $[n, k, 3]$ sobre $\mathbb{F}_4$ se essa sequência, ao ser devidamente mapeada para um vetor no espaço n-dimensional do $\mathbb{F}_4$, difere em até um símbolo de alguma palavra-código.

Além disso, avaliou-se a hipótese de existir um código BCH subjacente a uma sequência de DNA que revele semelhanças entre coleções de sequências de DNA de organismos de uma mesma classe taxonômica. Para isso, comparou-se a capacidade desses códigos de agrupar vetores distribuídos aleatoriamente em um espaço vetorial de dimensão n com a capacidade de agrupar sequências de DNA. Observou-se, por fim, que os códigos BCH que identificam a maioria das sequências de DNA de uma mesma coleção, não fornecem informações biológicas suficientes para permitir a classificação e agrupamento de tais sequências.

O Capítulo 5 foi dedicado à apresentação dos resultados referente à pesquisa da discriminação de sequências codificantes e não codificantes de proteínas por meio da análise espectral de tais sequências. Nesse contexto, o principal desafio é determinar qual a regra de mapeamento que resulta em uma representação numérica adequada para seguir com a análise espectral. Isto porque, ao comparar mapeamentos distintos para uma determinada sequência de DNA, os respectivos espectros de energia também são distintos e nem sempre representam versões aproximadas um do outro. Portanto, avaliou-se a hipótese de que, para esta aplicação, um mapeamento adaptativo deve ser mais adequado, assim, cada sequência deve ter uma regra de mapeamento particular, na qual, deve-se capturar o máximo possível de informações sobre essa sequência.

Foram propostos os algoritmos SNR-SE e TBP-SE para determinação de mapeamentos adaptativos, ambos baseados na envoltória espectral. Além disso, o uso prévio da propriedade TBP como critério de otimização foi útil no desenvolvimento do TBP-SE. A performance desses algoritmos foi avaliada comparando-os com outros algoritmos já consolidados na literatura, sendo eles: Voss [53], EIIP [54], QPSK [5], MEM [8]. Observou-se que o TBP-SE se destacou pois possui a maior acurácia e sensibilidade, ressaltando que, a taxa de sensibilidade é ainda mais importante nesta aplicação, pois, assim, reduzimos a probabilidade de ter uma sequência de codificação que não será identificada. Isso foi bem destacado ao considerar o caso específico do gene F56F11.4a, no qual só foi possível identificar a presença de todos os seus cinco éxons ao usar o mapeamento TBE-SE.

6.1 Trabalhos Futuros

A pesquisa sobre aplicabilidade de métodos de processamento da informação para o DNA é uma área ativa e ainda existem algumas questões que ainda não foram exploradas portanto, merecem atenção e continuidade da pesquisa, entre as quais:

- **Estruturas matemáticas para identificação do DNA**

  A comunidade científica vem contribuindo com algoritmos para tentar definir completamente um sequência de DNA a partir de estruturas algébricas. Além disso, alguns questionamentos adjacentes têm sido respondidos, assim como o que foi apresentado no Capítulo 4. Porém, a existência de uma estrutura algébrica bem como de um código corretor de erro capaz de identificar e reproduzir completamente sequências de DNA permanece sendo um problema em aberto. A relação entre os parâmetros do código (comprimento e estrutura algébrica) e de uma sequência de DNA ainda é limitada ao considerar a hipótese de que o código subjacente à essas sequências é o código BCH. Portanto, propõe-se investigar novas estruturas com o objetivo de contribuir com esse âmbito da pesquisa.

- **Identificação de periodicidades no DNA**

  Nesta pesquisa foram propostos dois algoritmos para representação numérica do DNA. Ambos os algoritmos eram baseados na envoltória espectral e resultavam em mapeamentos adaptativos para as sequências de DNA. A representação numérica do DNA por meio de mapeamento adaptativo demonstrou melhorias na discriminação entre sequências codificantes e não codificantes de proteína ao utilizar, como critério, a observação da propriedade TBP no espectro de energia dessas sequências. Sendo assim, propõe-se investigar se é possível discriminar outros tipos de regiões com frequências características, por exemplo, repetições tandem. As repetições tandem ocorrem no DNA quando um padrão de um ou mais nucleotídeos é repetido e as repetições são diretamente adjacentes umas às outras.

- **Potencial de discriminação de éxons com poucas bases**

  Um dos desafios da discriminação de regiões de codificação e não codificação do DNA é o tamanho da sequência. Alguns éxons são curtos e, por esta razão, aumenta a dificuldade de sua identificação. Na literatura existe o registro de éxons menores que 10 bases. Embora as análises experimentais realizadas no Capítulo 5 revelem que o algoritmo TBP-SE, proposto nesta Tese, tenha sido o único a identificar o primeiro éxon do gene F56F11.4a (em geral, esse é o éxon mais difícil de localizar pois é o mais curto desse gene), propõe-se ampliar a investigação da performance do método para localização desses éxons curtos.

Referências

1 DOUGHERTY, E. R.; SHMULEVICH, I. *Genomic signal processing and statistics*. EUA: Hindawi Publishing Corporation, 2005. v. 2. Citado na página 24.

2 KWAN, H. K.; ARNIKER, S. B. Numerical representation of DNA sequences. In: IEEE. *2009 IEEE International Conference on Electro/Information Technology*. Canada, 2009. p. 307–310. Citado na página 25.

3 MENDIZABAL-RUIZ, G.; ROMÁN-GODÍNEZ, I.; TORRES-RAMOS, S.; SALIDO-RUIZ, R. A.; MORALES, J. A. On DNA numerical representations for genomic similarity computation. *PloS one*, Public Library of Science San Francisco, CA USA, v. 12, n. 3, p. e0173288, 2017. Citado na página 25.

4 TIWARI, S.; RAMACHANDRAN, S.; BHATTACHARYA, A.; BHATTACHARYA, S.; RAMASWAMY, R. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics*, Oxford University Press, v. 13, n. 3, p. 263–270, 1997. Citado nas páginas 25, 28, 49, 51, 55 e 94.

5 ANASTASSIOU, D. Genomic signal processing. *IEEE Signal Processing Magazine*, v. 18, n. 4, p. 8–20, 2001. Citado nas páginas 25, 28, 29, 47, 49, 51, 52, 55, 81, 88, 89, 90, 93, 97 e 100.

6 YIN, C.; YAU, S. S.-T. A Fourier characteristic of coding sequences: origins and a non-Fourier approximation. *Journal of computational biology*, Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA, v. 12, n. 9, p. 1153–1165, 2005. Citado nas páginas 25, 51 e 84.

7 YIN, C.; YAU, S. S.-T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *Journal of theoretical biology*, Elsevier, v. 247, n. 4, p. 687–694, 2007. Citado nas páginas 25 e 51.

8 GALLEANI, L.; GARELLO, R. The minimum entropy mapping spectrum of a DNA sequence. *IEEE Transactions on Information Theory*, IEEE, v. 56, n. 2, p. 771–783, 2010. Citado nas páginas 25, 28, 29, 49, 51, 81, 87, 88, 89, 90, 93, 97 e 100.

9 SAHU, S. S.; PANDA, G. Identification of protein-coding regions in DNA sequences using a time-frequency filtering approach. *Genomics, proteomics & bioinformatics*, Elsevier, v. 9, n. 1-2, p. 45–55, 2011. Citado nas páginas 25, 28, 39, 51 e 52.

10 ROY, M.; BARMAN, S. Effective gene prediction by high resolution frequency estimator based on least-norm solution technique. *EURASIP Journal on Bioinformatics and Systems Biology*, Springer, v. 2014, n. 1, p. 2, 2014. Citado nas páginas 25, 28 e 51.

11 ROY, S. S.; BARMAN, S. Identification of protein coding region of DNA sequence using multirate filter. In: *Computational Advancement in Communication Circuits and Systems*. New Delhi: Springer, 2015. p. 131–137. ISBN 978-81-322-2274-3. Citado nas páginas 25, 51 e 52.

12 ADALBJORNSSON, S. I.; SWARD, J.; WALLIN, J.; JAKOBSSON, A. Estimating periodicities in symbolic sequences using sparse modeling. *IEEE Transactions on Signal Processing*, IEEE, v. 63, n. 8, p. 2142–2150, 2015. Citado nas páginas 25 e 51.

13 LI, J.; ZHANG, L.; LI, H.; PING, Y.; XU, Q.; WANG, R.; TAN, R.; WANG, Z.; LIU, B.; WANG, Y. Integrated entropy-based approach for analyzing exons and introns in DNA sequences. *BMC bioinformatics*, BioMed Central, v. 20, n. 8, p. 1–7, 2019. Citado nas páginas 25 e 51.

14 SINGH, A. K.; SRIVASTAVA, V. K. The Three Base Periodicity of Protein Coding Sequences and its Application in Exon Prediction. In: IEEE. *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. India, 2020. p. 1089–1094. Citado nas páginas 25, 49, 51, 52 e 94.

15 GABRIELIAN, A.; PONGOR, S. Correlation of intrinsic DNA curvature with DNA property periodicity. *FEBS letters*, Wiley Online Library, v. 393, n. 1, p. 65–68, 1996. Citado na página 25.

16 SHARMA, D.; ISSAC, B.; RAGHAVA, G.; RAMASWAMY, R. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*, Oxford University Press, v. 20, n. 9, p. 1405–1412, 2004. Citado na página 25.

17 PAUL, T.; VAINIO, S.; RONING, J. Haar wavelet based approach for Short Tandem Repeats (STR) Detection. In: IEEE. *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. Emirados Árabes Unidos, 2019. p. 1–6. Citado na página 25.

18 CHEEVER, E.; SEARLS, D.; KARUNARATNE, W.; OVERTON, G. Using signal processing techniques for DNA sequence comparison. In: IEEE. *Proceedings of the Fifteenth Annual Northeast Bioengineering Conference*. USA, 1989. p. 173–174. Citado na página 25.

19 BORRAYO, E.; MENDIZABAL-RUIZ, E. G.; VÉLEZ-PÉREZ, H.; ROMO-VÁZQUEZ, R.; MENDIZABAL, A. P.; MORALES, J. A. Genomic signal processing methods for computation of alignment-free distances from DNA sequences. *PloS one*, Public Library of Science, v. 9, n. 11, p. e110954, 2014. Citado na página 25.

20 KING, B. R.; ABURDENE, M.; THOMPSON, A.; WARRES, Z. Application of discrete Fourier inter-coefficient difference for assessing genetic sequence similarity. *EURASIP Journal on Bioinformatics and Systems Biology*, Springer, v. 2014, n. 1, p. 8, 2014. Citado na página 25.

21 SKUTKOVA, H.; VITEK, M.; SEDLAR, K.; PROVAZNIK, I. Progressive alignment of genomic signals by multiple dynamic time warping. *Journal of theoretical biology*, Elsevier, v. 385, p. 20–30, 2015. Citado na página 25.

22 JAYAPRIYA, J.; AROCK, M. Aligning molecular sequences by wavelet transform using cross correlation similarity metric. *Int. J. Intell. Syst. Appl.(IJISA)*, v. 9, n. 11, p. 62–70, 2017. Citado na página 25.

23 OTU, H. H.; SAYOOD, K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, Oxford University Press, v. 19, n. 16, p. 2122–2130, 2003. Citado nas páginas 25 e 60.

24 LI, B.; LI, Y.-B.; HE, H.-B. LZ complexity distance of DNA sequences and its application in phylogenetic tree reconstruction. *Genomics, proteomics & bioinformatics*, Elsevier, v. 3, n. 4, p. 206–212, 2005. Citado nas páginas 25 e 60.

25 LIU, N.; WANG, T.-m. A relative similarity measure for the similarity analysis of DNA sequences. *Chemical Physics Letters*, Elsevier, v. 408, n. 4-6, p. 307–311, 2005. Citado na página 25.

26 HOANG, T.; YIN, C.; YAU, S. S.-T. Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics*, v. 108, n. 3, p. 134 – 142, 2016. ISSN 0888-7543. Citado nas páginas 25, 26 e 52.

27 SOLIS-REYES, S.; AVINO, M.; POON, A.; KARI, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS One*, Public Library of Science San Francisco, CA USA, v. 13, n. 11, p. e0206409, 2018. Citado nas páginas 25, 26 e 80.

28 HE, L.; DONG, R.; HE, R. L.; YAU, S. S.-T. A novel alignment-free method for HIV-1 subtype classification. *Infection, Genetics and Evolution*, Elsevier, v. 77, p. 104080, 2020. Citado nas páginas 25 e 26.

29 CHURCH, G. M.; GAO, Y.; KOSURI, S. Next-generation digital information storage in DNA. *Science*, American Association for the Advancement of Science, v. 337, n. 6102, p. 1628–1628, 2012. ISSN 0036-8075. Citado na página 25.

30 ERLICH, Y.; ZIELINSKI, D. DNA Fountain enables a robust and efficient storage architecture. *Science (New York, N.Y.)*, American Association for the Advancement of Science, v. 355, n. 6328, p. 950–954, 2017. ISSN 1095-9203. Citado na página 25.

31 CLELLAND, C. T.; RISCA, V.; BANCROFT, C. Hiding messages in DNA microdots. *Nature*, Nature Publishing Group, v. 399, n. 6736, p. 533–534, 1999. ISSN 0028-0836. Citado na página 25.

32 BECK, M.; YAMPOLSKIY, R. DNA as a medium for hiding data. *BMC bioinformatics*, v. 13, n. 12, p. 1–1, 2012. Citado na página 25.

33 RODRIGUEZ-SARMIENTO, D. L.; DUARTE-GONZALEZ, M. E.; RODRIGUEZ-QUINONES, T.; PALAZZO, R. Procedure for identifying odd-sized nucleotide sequences as codewords of BCH codes over GF(4). In: IEEE. *2019 IEEE International Symposium on Information Theory (ISIT)*. France, 2019. p. 922–926. Citado nas páginas 25, 26, 75 e 79.

34 HERNADEZ, G. L.; DUARTE-GONZALEZ, M. E.; PALAZZO, R. Identification of odd-sized DNA and mRNA sequences as codewords of BCH codes over Z4. In: IEEE. *2019 IEEE International Conference on Applied Science and Advanced Technology (iCASAT)*. México, 2019. p. 1–6. Citado nas páginas 25 e 26.

35 SCHNEIDER, T. D.; STORMO, G. D.; GOLD, L.; EHRENFEUCHT, A. Information content of binding sites on nucleotide sequences. *Journal of Molecular Biology*, v. 188, n. 3, p. 415 – 431, 1986. ISSN 0022-2836. Citado na página 25.

36 YOCKEY, H. P. *Information theory and molecular biology*. Reino Unido: Cambridge University Press, 1992. ISBN 9780521359030. Citado na página 25.

37 BATTAIL, G. Information Theory and Error-Correcting Codes In Genetics and Biological Evolution. In: *Introduction to Biosemiotics: The New Biological Synthesis*. Dordrecht: Springer Netherlands, 2007. p. 299–345. Citado na página 25.

38 FARIA, L.; ROCHA, A.; PALAZZO, R. Transmission of intra-cellular genetic information: A system proposal. *Journal of theoretical biology*, Elsevier, v. 358, p. 208–231, 2014. Citado na página 25.

39 DUARTE-GONZáLEZ, M. E.; ECHEVERRI, O. Y.; GUEVARA, J. M.; PALAZZO, R. Cyclic Concatenated Genetic Encoder: A mathematical proposal for biological inferences. *Biosystems*, Elsevier, v. 163, p. 47–58, 2018. Citado nas páginas 25 e 26.

40 LIEBOVITCH, L. S.; TAO, Y.; TODOROV, A. T.; LEVINE, L. Is there an error correcting code in the base sequence in DNA? *Biophysical Journal*, Elsevier, v. 71, n. 3, p. 1539–1544, 1996. Citado na página 25.

41 ROSEN, G. Examining coding structure and redundancy in DNA. *IEEE engineering in medicine and biology magazine*, IEEE, v. 25, n. 1, p. 62–68, 2006. Citado na página 25.

42 FARIA, L.; ROCHA, A.; KLEINSCHMIDT, J.; PALAZZO, R.; SILVA-FILHO, M. DNA sequences generated by BCH codes over GF (4). *Electronics letters*, IET, v. 46, n. 3, p. 203–204, 2010. Citado nas páginas 25, 67, 68, 74, 75 e 79.

43 ROCHA, A. S. L.; FARIA, L. C. B.; KLEINSCHMIDT, J. H.; PALAZZO, R.; SILVA-FILHO, M. C. DNA sequences generated by Z4-linear codes. In: IEEE. *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. USA, 2010. p. 1320–1324. Citado nas páginas 25 e 67.

44 FARIA, L.; ROCHA, A.; KLEINSCHMIDT, J.; SILVA-FILHO, M.; BIM, E.; HERAI, R. H.; YAMAGISHI, M. E.; PALAZZO, R. Is a genome a codeword of an error-correcting code? *PloS one*, Public Library of Science, v. 7, n. 5, p. e36644, 2012. Citado nas páginas 25 e 26.

45 BRANDÃO, M. M.; SPOLADORE, L.; FARIA, L. C.; ROCHA, A. S.; SILVA-FILHO, M. C.; PALAZZO, R. Ancient DNA sequence revealed by error-correcting codes. *Scientific reports*, Nature Publishing Group, v. 5, p. 12051, 2015. Citado na página 26.

46 TRIFONOV, E. N.; SUSSMAN, J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 77, n. 7, p. 3816–3820, 1980. Citado na página 26.

47 FICKETT, J. W. Recognition of protein coding regions in DNA sequences. *Nucleic acids research*, Oxford University Press, v. 10, n. 17, p. 5303–5318, 1982. Citado nas páginas 26 e 51.

48 SHEPHERD, J. C. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. *Journal of Molecular Evolution*, Springer, v. 17, n. 2, p. 94–102, 1981. Citado na página 27.

49 SHEPHERD, J. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 78, n. 3, p. 1596–1600, 1981. Citado na página 27.

50 TSONIS, A. A.; ELSNER, J. B.; TSONIS, P. A. Periodicity in DNA coding sequences: implications in gene evolution. *Journal of theoretical biology*, Elsevier, v. 151, n. 3, p. 323–331, 1991. Citado na página 27.

51 SÁNCHEZ, J.; LOPEZ-VILLASENOR, I. A simple model to explain three-base periodicity in coding DNA. *FEBS letters*, Elsevier, v. 580, n. 27, p. 6413–6422, 2006. Citado na página 27.

52 HOWE, E. D.; SONG, J. S. Categorical spectral analysis of periodicity in human and viral genomes. *Nucleic acids research*, Oxford University Press, v. 41, n. 3, p. 1395–1405, 2013. Citado na página 27.

53 VOSS, R. F. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical review letters*, APS, v. 68, n. 25, p. 3805, 1992. Citado nas páginas 27, 29, 48, 81, 88, 90, 93, 97 e 100.

54 LALOVIC, D.; VELJKOVIC, V. The global average DNA base composition of coding regions may be determined by the electron-ion interaction potential. *Biosystems*, v. 23, n. 4, p. 311–316, 1990. ISSN 0303-2647. Citado nas páginas 28, 48, 81, 88, 90, 93, 97 e 100.

55 VAIDYANATHAN, P.; YOON, B.-J. Digital filters for gene prediction applications. In: IEEE. *Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002.* USA, 2002. v. 1, p. 306–310. Citado nas páginas 28, 39 e 56.

56 WANG, W.; JOHNSON, D. Computing linear transforms of symbolic signals. *IEEE Transactions on Signal Processing*, v. 50, n. 3, p. 628–634, Aug 2002. Citado nas páginas 28, 58, 86 e 96.

57 STOFFER, D. S.; TYLER, D. E.; MCDOUGALL, A. J. Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, Oxford University Press, v. 80, n. 3, p. 611–622, Sep 1993. Citado nas páginas 28, 58 e 96.

58 LALOVIC, D.; VELJKOVIC, V. The global average DNA base composition of coding regions may be determined by the electron-ion interaction potential. *Biosystems*, v. 23, n. 4, p. 311–316, 1990. ISSN 0303-2647. Citado na página 29.

59 EDWARDS, D.; HANSEN, D.; STAJICH, J. E. DNA Sequence Databases. In: *Bioinformatics*. New York: Springer New York, 2009. p. 1–11. Citado na página 29.

60 NCBI. *Nucleotide[Internet]*. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information. 1988. Disponível em: <https://www.ncbi.nlm.nih.gov/nuccore>. Acesso em: 17 sep 2020. Citado na página 29.

61 ARRUDA, M.; SILVA, A.; ASSIS, F. M. An Adaptive Mapping Method Using Spectral Envelope Approach for DNA Spectral Analysis. *Entropy*, v. 24, n. 7, 2022. ISSN 1099-4300. Citado nas páginas 30 e 81.

62 ARRUDA, M. M.; SILVA, A. da; ASSIS, F. M. Maximizing the SNR of DNA Spectrum for Coding Sequence Identification. In: *Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. Brasil: Sociedade Brasileira de Telecomunicações, 2021. Citado na página 30.

63 SILVA, A. da; ARRUDA, M. M.; ASSIS, F. M. Reconstrução de Árvores Filogenéticas a partir de mtDNA usando o Algoritmo SEQUITUR. In: *Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*. Brasil: Sociedade Brasileira de Telecomunicações, 2021. Citado nas páginas 30 e 61.

64 ARRUDA, M. M.; ASSIS, F. M.; SOUZA, T. A. Is BCH Code Useful to DNA Classification as an Alignment-Free Method? *IEEE Access*, v. 9, p. 68552–68560, 2021. ISSN 2169-3536. Citado nas páginas 30 e 66.

65 SHANNON, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, v. 27, p. 379–423, july, october 1948. Citado na página 33.

66 FANO, R. *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: The MIT Press, 1961. Citado na página 34.

67 HOCQUENGHEM, A. Codes correcteurs d'erreurs. *Chiffres*, v. 2, n. 2, p. 147–56, 1959. Citado na página 35.

68 BOSE, R. C.; RAY-CHAUDHURI, D. K. On a Class of Error Correcting Binary Group Codes. *Information and control*, Academic Press, v. 3, n. 1, p. 68–79, 1960. Citado na página 35.

69 BLAHUT, R. E. *Algebraic Codes for Data Transmission*. USA: Cambridge University Press, 2003. Citado nas páginas 36, 75, 76 e 119.

70 OPPENHEIM, A.; WILLSKY, A.; NAWAB, S.; HAMID, W.; YOUNG, I. *Signals & Systems*. USA: Prentice Hall, 1997. (Prentice-Hall signal processing series). ISBN 9780138147570. Citado na página 38.

71 CARVALHO, J.; GURJAO, E.; VELOSO, L. *Introdução à análise de sinais e sistemas*. Brasil: Elsevier, 2015. ISBN 9788535282368. Citado na página 38.

72 SEJDIĆ, E.; DJUROVIĆ, I.; JIANG, J. Time-frequency feature representation using energy concentration: An overview of recent advances. *Digital signal processing*, Elsevier, v. 19, n. 1, p. 153–183, 2009. Citado na página 39.

73 SMITHIII, J. O. *Spectral audio signal processing*. USA: W3K publishing, 2011. Citado na página 39.

74 PROAKIS, J. G. *Digital signal processing: principles algorithms and applications*. 3. ed. India: Pearson Education India, 2001. Citado na página 39.

75 SOLOMONOFF, R. J. A formal theory of inductive inference. Part I. *Information and control*, Elsevier, v. 7, n. 1, p. 1–22, 1964. Citado na página 41.

76 KOLMOGOROV, A. N. Three approaches to the quantitative definition of information. *Problems of information transmission*, v. 1, n. 1, p. 1–7, 1965. Citado na página 41.

77 COVER, T. M.; THOMAS, J. A. *Elements of information theory*. USA: Wiley Interscience, 2006. ISBN 9780471241959. Citado na página 41.

78 LEMPEL, A.; ZIV, J. On the complexity of finite sequences. *IEEE Transactions on information theory*, IEEE, v. 22, n. 1, p. 75–81, 1976. Citado nas páginas 42, 43, 44 e 60.

79 ZIV, J.; LEMPEL, A. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, v. 23, n. 3, p. 337–343, 1977. Citado na página 42.

80 ZIV, J.; LEMPEL, A. Compression of individual sequences via variable-rate coding. *IEEE transactions on Information Theory*, IEEE, v. 24, n. 5, p. 530–536, 1978. Citado na página 42.

81 WELCH, T. A. A technique for high-performance data compression. *Computer*, IEEE, n. 6, p. 8–19, 1984. Citado na página 42.

82 CRISTEA, P. D. Conversion of nucleotides sequences into genomic signals. *Journal of cellular and molecular medicine*, Wiley Online Library, v. 6, n. 2, p. 279–303, 2002. Citado na página 47.

83 CHAKRAVARTHY, N.; SPANIAS, A.; IASEMIDIS, L. D.; TSAKALIS, K. Autoregressive modeling and feature analysis of DNA sequences. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2004, n. 1, p. 952689, 2004. Citado nas páginas 47 e 48.

84 SILVERMAN, B.; LINSKER, R. A measure of DNA periodicity. *Journal of theoretical biology*, v. 118, n. 3, p. 295, 1986. Citado nas páginas 48 e 51.

85 JEFFREY, H. J. Chaos game representation of gene structure. *Nucleic acids research*, Oxford University Press, v. 18, n. 8, p. 2163–2170, 1990. Citado na página 48.

86 YU, X.; MEI, Z.; CHEN, C.; CHEN, W. Ranking Power Spectra: A Proof of Concept. *Entropy*, v. 21, n. 11, 2019. ISSN 1099-4300. Citado nas páginas 56, 87 e 93.

87 OYELADE, J.; ISEWON, I.; OLADIPUPO, F.; AROMOLARAN, O.; UWOGHIREN, E.; AMEH, F.; ACHAS, M.; ADEBIYI, E. Clustering algorithms: their application to gene expression data. *Bioinformatics and Biology insights*, SAGE Publications Sage UK: London, England, v. 10, p. BBI–S38316, 2016. Citado na página 59.

88 CHEN, X.; KWONG, S.; LI, M. A compression algorithm for dna sequences and its applications in genome comparison. In: *Proceedings of the fourth annual international conference on Computational molecular biology*. New York, NY, USA: Association for Computing Machinery, 2000. p. 107. Citado na página 59.

89 LI, M.; BADGER, J. H.; CHEN, X.; KWONG, S.; KEARNEY, P.; ZHANG, H. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, Oxford University Press, v. 17, n. 2, p. 149–154, 2001. Citado na página 60.

90 LIU, L.; LI, D.; BAI, F. A relative Lempel–Ziv complexity: Application to comparing biological sequences. *Chemical Physics Letters*, Elsevier, v. 530, p. 107–112, 2012. Citado na página 60.

91 SENGUPTA, D. C.; HILL, M. D.; BENTON, K. R.; BANERJEE, H. N. Similarity studies of corona viruses through chaos game representation. *Computational molecular bioscience*, NIH Public Access, v. 10, n. 3, p. 61, 2020. Citado na página 60.

92 NEVILL-MANNING, C. G.; WITTEN, I. H.; MAULSBY, D. L. Compression by induction of hierarchical grammars. In: IEEE. *Proceedings of IEEE Data Compression Conference (DCC'94)*. USA, 1994. p. 244–253. Citado na página 61.

93 NEVILL-MANNING, C. G.; WITTEN, I. H. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, AI Access Foundation, v. 7, p. 67–82, 1997. Citado na página 61.

94 FOLEY, N. M.; SPRINGER, M. S.; TEELING, E. C. Mammal madness: is the mammal tree of life not yet resolved? *Philosophical Transactions of the Royal Society B: Biological Sciences*, The Royal Society, v. 371, n. 1699, p. 20150140, 2016. Citado na página 64.

95 NCBI. *Genome Data Viewer*. Bethesda (MD): National Center for Biotechnology Information (US). 1998. Disponível em: <https://www.ncbi.nlm.nih.gov/genome/gdv/>. Citado na página 64.

96 ARRUDA, M. *Milena-Arruda/DNA-BCH*. Disponível em: <https://github.com/Milena-Arruda/dna-bch>. Citado na página 66.

97 PAO, S. S.; PAULSEN, I. T.; SAIER, M. H. Major facilitator superfamily. *Microbiology and molecular biology reviews*, Am Soc Microbiol, v. 62, n. 1, p. 1–34, 1998. Citado na página 73.

98 ARRUDA, M. *Milena-Arruda/dna-spectral-analysis*. Disponível em: <https://github.com/Milena-Arruda/dna-spectral-analysis>. Citado na página 81.

99 NIELSEN, M. A.; CHUANG, I. *Quantum computation and quantum information*. USA: Cambridge University Press, New York, 2001. Citado na página 88.

100 PUTLURI, S.; RAHMAN, M. Z. U.; AMARA, C. S.; PUTLURI, N. New exon prediction techniques using adaptive signal processing algorithms for genomic analysis. *IEEE Access*, IEEE, v. 7, p. 80800–80812, 2019. Citado nas páginas 89 e 94.

101 SETUBAL, J. C.; MEIDANIS, J. *Introduction to computational molecular biology*. USA: PWS Pub. Boston, 1997. Citado nas páginas 112 e 131.

102 JONES, N. C.; PEVZNER, P. A. *An introduction to bioinformatics algorithms*. USA: MIT press, 2004. Citado na página 112.

103 COMPEAU, P.; PEVZNER, P. A. *Bioinformatics Algorithms: An Active Learning Approach. La Jolla*. USA: CA: Active Learning Publishers, 2018. Citado na página 112.

104 ORGEL, L. E. The Origin of Life on the Earth. *Scientific American*, JSTOR, v. 271, n. 4, p. 76–83, 1994. Citado na página 112.

105 NCBI. *Genes and Disease [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US). 1998. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK22183/>. Citado na página 115.

106 HEFEZ, A.; VILLELA, M. L. T. *Códigos Corretores de Erros*. Brasil: Instituto Nacional de Matemática Pura e Aplicada, 2008. Citado na página 119.

107 SOKAL, R. R. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, v. 38, p. 1409–1438, 1958. Citado na página 133.

Apêndices

APÊNDICE A

Conceitos Básicos da Biologia Molecular

Neste capítulo serão apresentados conceitos básicos da biologia molecular que são importantes para acompanhar o *background* biológico desta proposta de tese. Portanto, serão descritos a estrutura básica e a função das proteínas e dos ácidos nucleicos, os mecanismos da genética molecular e uma visão geral dos bancos de dados de sequências existentes. As definições tratadas ao longo deste capítulo foram extraídas de [101, 102, 103].

A vida

A ciência moderna mostrou que a vida começou há cerca de 3,5 bilhões de anos. As semelhanças entre organismos vivos indicam a presença de um ancestral comum a partir do qual todas as espécies divergiram por um processo de evolução. Os cientistas argumentam que esse ancestral comum existe pois seria impossível que tais traços universais tenham evoluído separadamente. Para justificar essa hipótese, Orgel [104] faz uma analogia com dois roteiros virtualmente idênticos (*scripts*), diferindo apenas em poucas palavras. Ele aponta que não seria razoável pensar que os *scripts* foram criados independentemente por dois autores separados. Porém, seria seguro assumir que um *script* era uma réplica imperfeita do outro ou que ambas as versões eram cópias ligeiramente alteradas de um terceiro.

A biologia ao nível microscópico teve seu início em 1665 quando Robert Hooke, ao observar cortes de cortiça (material de origem vegetal utilizado para fazer rolhas), descobriu que os organismos são compostos de células. Em seguida, na década de 1830, Matthias Schleiden e Theodor Schwann propuseram a teoria celular que transformou a biologia em uma ciência além do alcance do olho nu. De muitas maneiras, o estudo da vida tornou-se o estudo das células.

Todos os organismos vivos, sejam eles simples ou complexos, apesentam uma

bioquímica molecular semelhante, cujas principais estruturas são as proteínas e os ácidos nucleicos. De modo geral, as proteínas são responsáveis por aquilo que um ser vivo é e faz no sentido físico, e os ácidos nucleicos, por outro lado, codificam a informação necessária para produzir as proteínas e são responsáveis por repassar essas informações para as gerações subsequentes.

Proteínas

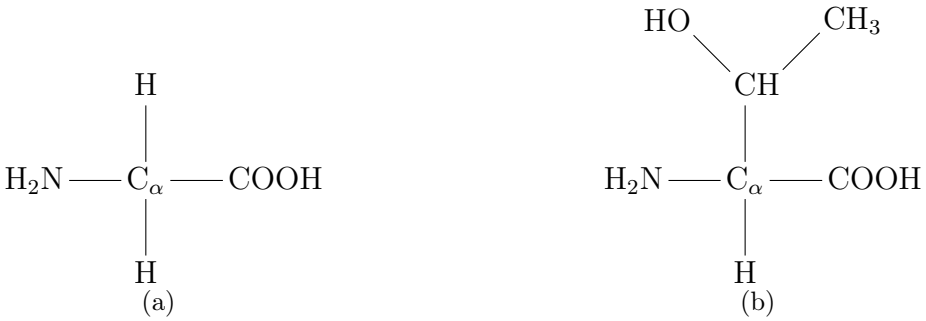
As proteínas são macromoléculas biológicas constituídas por cadeias de aminoácidos, cuja estrutura tridimensional determina a função que a mesma desempenha. Por exemplo, a estrutura de uma proteína pode ser capaz de se ligar a várias cópias idênticas de si mesma, construindo um fio de cabelo. Ou a estrutura pode ligar moléculas diferentes à proteína e começam a trocar átomos, assim, a proteína está cumprindo seu papel como catalisador.

Em 1820, Henry Braconnot identificou o primeiro aminoácido, a glicina, e no início dos anos de 1900, todos os vinte aminoácidos especificados pelo código genético já haviam sido descobertos e sua estrutura química identificada. Os aminoácidos são compostos quaternários com um átomo de carbono central conhecido como carbono α, que está ligado a um átomo de hidrogênio, um grupo amina (NH_2), um grupo carboxila (COOH) e uma cadeia lateral (responsável por distinguir um aminoácido do outro), conforme os exemplos da Figura A.1.

Ácidos Nucleicos

Os ácidos nucleicos presentes nos organismos vivos são: ácido desoxirribonucleico (DNA, do inglês: *deoxyribonucleic acid*) e ácido ribonucleico (RNA, do inglês: *ribonucleic acid*). O DNA foi descoberto em 1869 por Johann Friedrich Miescher ao isolar uma substância do núcleo dos glóbulos brancos. Na década de 1920, os ácidos nucleicos foram agrupados em duas classes chamadas DNA e RNA cuja principal diferença é o açúcar e a

Figura A.1 – Exemplos de aminoácidos: (a) glicina (Gly / G); (b) treonina (Thr / T).



Fonte: Elaborada pela autora.

composição de base. Ambas são moléculas simples que consistem de um açúcar (para o DNA têm-se 2′-desoxirribose e para o RNA têm-se ribose), um grupo fosfato e uma das quatro bases nitrogenadas (adenina (`A`), citosina (`C`), guanina (`G`) ou timina (`T`) para DNA ou uracila (U) para RNA), conforme ilustrado na Figura A.2. As ligações químicas que unem os nucleotídeos no DNA são sempre as mesmas, de modo que a estrutura de uma molécula de DNA é muito regular e são as bases que dão individualidade a cada molécula de DNA.

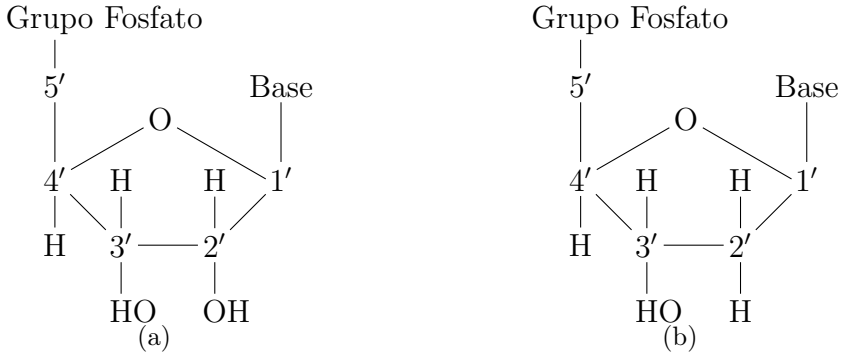
A estrutura em dupla hélice do DNA foi modelada em 1953 por James Watson e Francis Crick. A conexão da dupla cadeia dar-se-á por meio das bases, sendo `A` e `C` sempre emparelhadas com `T` e `G`, respectivamente, esses arranjos são chamados pares de bases complementares. Na Figura A.3 têm-se um esquemático da estrutura química do DNA indicando suas bases. Por convenção, a orientação da molécula de DNA começa na extremidade 5′ e termina na extremidade 3′, e toda a informação contida em uma fita de DNA está também contida na outra a partir da operação complementação reversa, conforme Exemplo A.1.

***Exemplo* A.1** *Considerando a sequência* $s =$ `GTGACCCTGGCCAGGACTGAC` *na direção canônica. Na direção reversa obtém-se* $s' =$ `CAGTCAGGACCGGTCCCAGTG`. *O complemento das bases de* s' *corresponde a operação de complementação reversa e resulta em* $\overline{s} =$ `GTCAGTCCTGGCCAGGGTCAC`. *Essa sequência tem comprimento 21.*

Código Genético

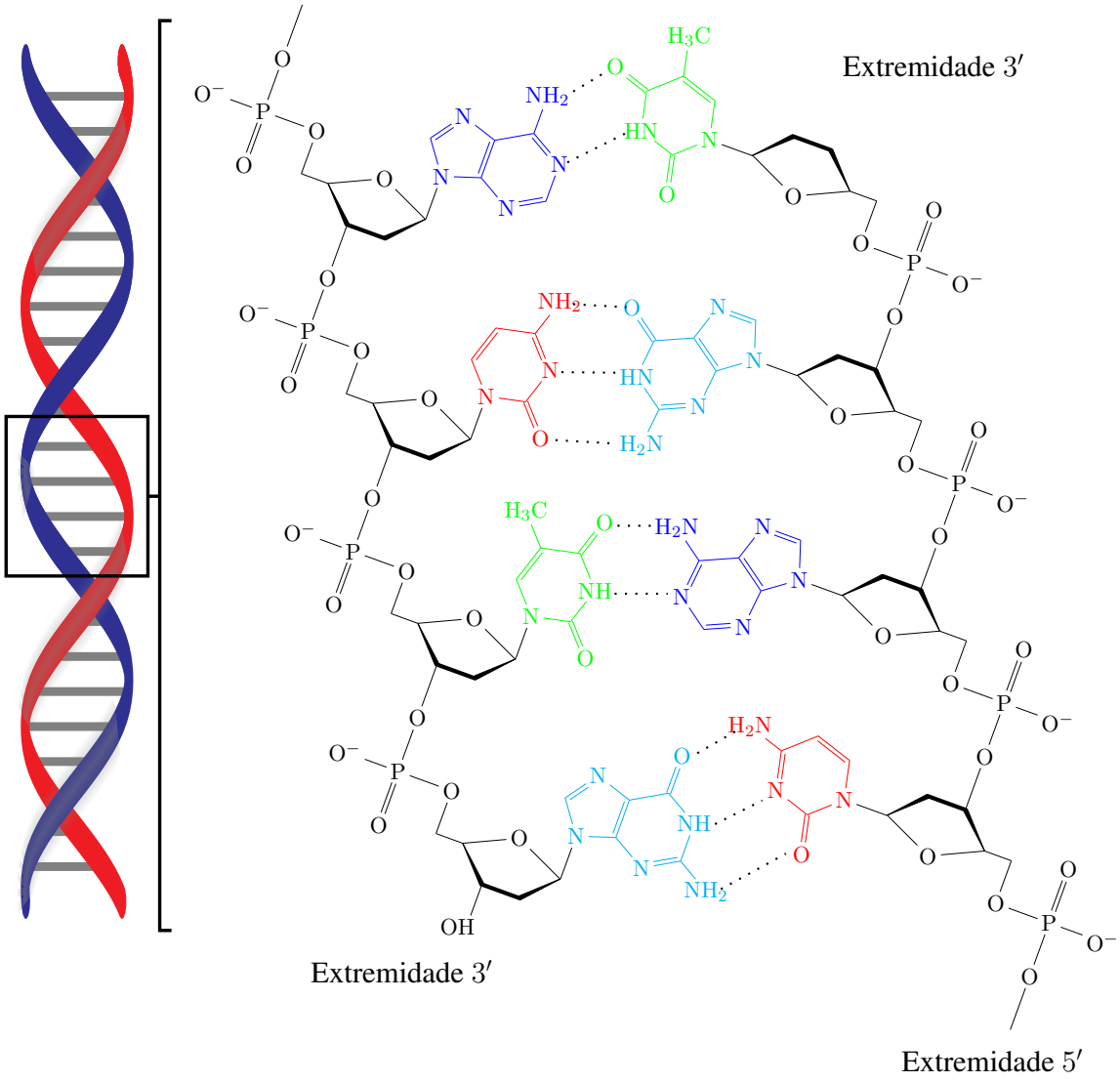
A informação presente no DNA, ou informação genética, é fundamental para construção de cada proteína ou RNA encontrado em um organismo. As células que encapsulam seu DNA em um núcleo são referidas como células eucarióticas; e aquelas cujo DNA está livre são células procarióticas. Todos os organismos multicelulares (como

Figura A.2 – Ácidos nucleicos: (a) a ribose está presente no RNA; (b) o 2′-desoxirribose está presente no DNA; cuja diferença é o oxigênio no carbono 2′. Os símbolos 1′ a 5′ representam átomos de carbono.



Fonte: Elaborada pela autora.

Figura A.3 – Esquemático da estrutura química do DNA com as quatro bases: adenina (em azul), citosina (em vermelho), guanina (em ciano) e timina (em verde).



Fonte: Elaborada pela autora.

pássaros ou humanos) são eucarióticos, enquanto a maioria dos organismos unicelulares (como bactérias) são procarióticos.

Os genes correspondem aos trechos contínuos do DNA que codificam informações para a construção de proteínas. Para nossos propósitos, a principal diferença entre procariotos e eucariotos é que os genes procariotos são cadeias contínuas, enquanto os eucariotos são quebrados em pedaços, chamados éxons. As cadeias não traduzidas, que intercalam os éxons, são os íntrons. Por exemplo, os genes humanos podem ser divididos em até 50 éxons, que representam menos de 5% do genoma (a função do DNA remanescente não é clara) e alguns cromossomos possuem uma densidade de genes maior do que outros [105].

Para especificar uma proteína é necessário identificar cada aminoácido que a contém. Isto é feito a partir dos códons, triplas de nucleotídeos do RNA. Porém, existem $4^3 = 64$ diferentes códons, que é mais de três vezes maior que o número de aminoácidos

Tabela A.1 – Código genético que mapeia códons para aminoácidos. O aminoácido methionine (AUG) também atua como códon de iniciação na transcrição.

Primeira Base	Segunda Base: G		Segunda Base: A		Segunda Base: C		Segunda Base: U		Terceira Base
G	GGG	Glycine (Gly/G)	GAG	Glutamic (Glu/E)	GCG	Alanine (Ala/A)	GUG	Valine (Val/V)	G
	GGA		GAA		GCA		GUA		A
	GGC		GAC	Aspartic (Asp/D)	GCC		GUC		C
	GGU		GAU		GCU		GUU		U
A	AGG	Arginine (Arg/R)	AAG	Lysine (Lys/K)	ACG	Threonine (Thr/T)	AUG	Methionine	G
	AGA		AAA		ACA		AUA	Isoleucine (Ile/I)	A
	AGC	Serine (Ser/S)	AAC	Asparagine (Asn/N)	ACC		AUC		C
	AGU		AAU		ACU		AUU		U
C	CGG	Arginine (Arg/R)	CAG	Glutamine (Gln/Q)	CCG	Proline (Pro/P)	CUG	Leucine (Leu/L)	G
	CGA		CAA		CCA		CUA		A
	CGC		CAC	Histidine (His/H)	CCC		CUC		C
	CGU		CAU		CCU		CUU		U
U	UGG	Tryptophan	UAG	Stop	UCG	Serine (Ser/S)	UUG	Leucine (Leu/L)	G
	UGA	Stop	UAA		UCA		UUA		A
	UGC	Cysteine (Cys/C)	UAC	Tyrosine (Tyr/Y)	UCC		UUC	Phenylalanine (Phe/F)	C
	UGU		UAU		UCU		UUU		U

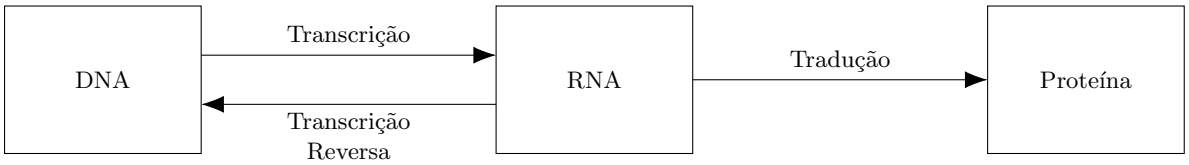
especificados. Atribui-se a existência dessa redundância à degeneração do código genético. No final dos anos de 1960, os cientistas mapearam todo o código genético conforme a Tabela A.1.

Síntese Proteica

O fluxo de informação genética dentro de um sistema biológico é referido como o dogma central da biologia molecular. A Figura A.4 resume os processos que incluem o dogma central. Muitos sistemas químicos na célula requerem enzimas (proteínas que atuam como catalisadores biológicos), é o caso da síntese de proteínas que pode ser dividida amplamente em duas fases: transcrição e tradução.

A transcrição é realizada por enzimas, conhecidas como RNA polimerases, no núcleo da célula. Durante esse processo, uma seção de DNA que codifica uma proteína é copiada em uma molécula chamada RNA mensageiro (por exemplo, pareando um DNA T com um RNA A, um DNA A com um RNA U e assim por diante). Em eucariotos, este RNA mensageiro (mRNA) é inicialmente produzido em uma forma prematura (pré-mRNA) que sofre um processamento, incluindo o *splicing*, para produzir mRNA maduro. Durante o *splicing* os íntrons são removidos e os éxons são unidos.

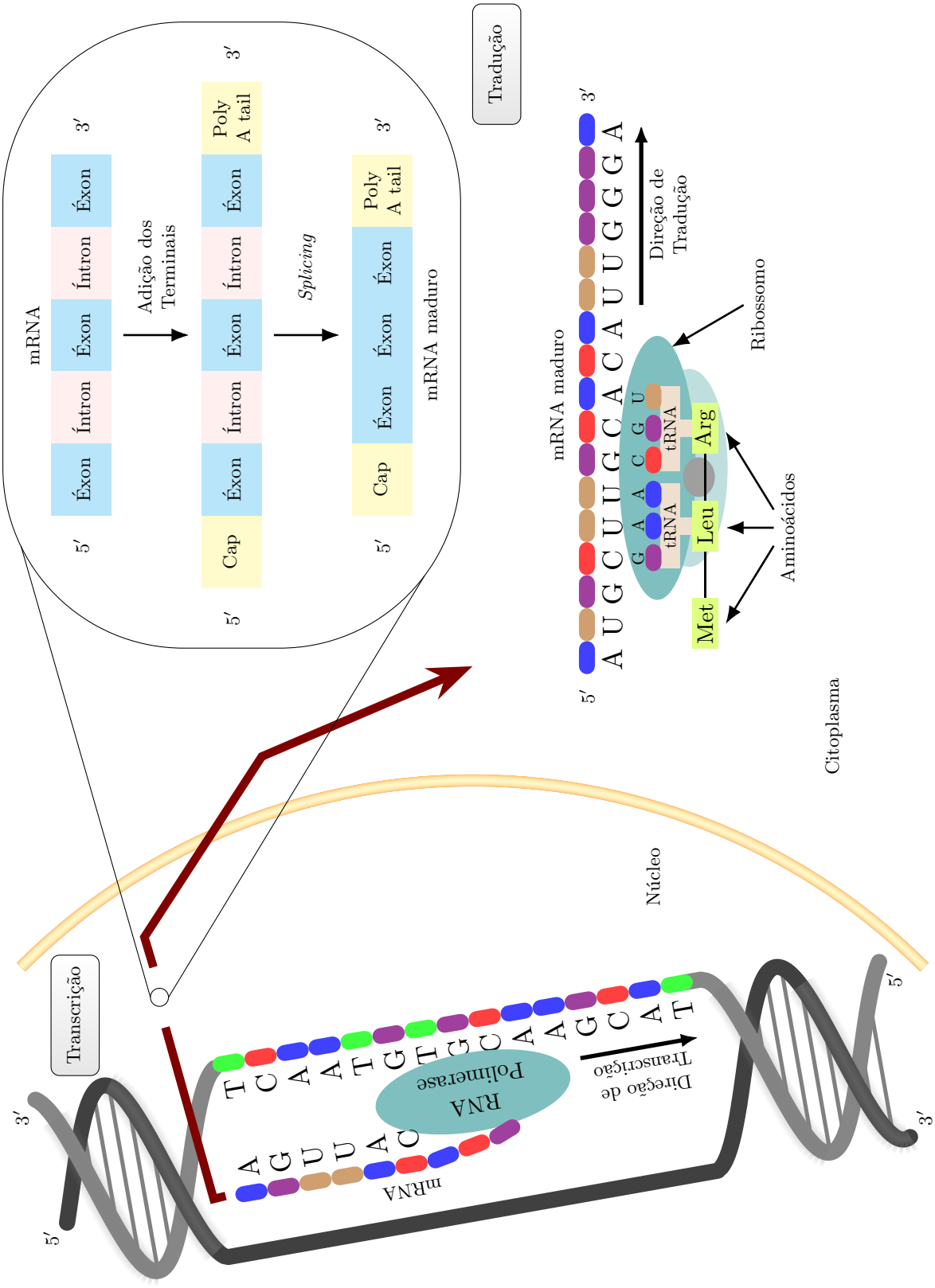
Figura A.4 – Fluxo de informações genéticas em uma célula: dogma central da biologia.



Fonte: Elaborada pela autora.

O mRNA maduro é exportado do núcleo, por meio de poros nucleares, para o citoplasma da célula para que ocorra a tradução. Essa molécula é então atacada por grandes complexos moleculares conhecidos como ribossomos, que leem códons consecutivos e localizam o aminoácido correspondente para inclusão na crescente cadeia polipeptídica. Os aminoácidos são localizados por meio de um tipo especial de RNA, denominado RNA de transferência (tRNA). Existem vinte tipos de tRNAs e vinte tipos de aminoácidos, cada tipo de aminoácido se liga a um tRNA diferente. As moléculas de tRNA têm um segmento de três bases (denominado anticódon) que é complementar ao códon no mRNA e adere ao mesmo, o que torna o aminoácido disponível para o ribossomo adicionar à cadeia polipeptídica. Quando um aminoácido é adicionado, o ribossomo desloca um códon para a direita e o processo se repete. É através desse processo que todas as proteínas são produzidas, incluindo as que são necessárias para realizá-lo. Um esquemático de todo esse processo é exposto na Figura A.5.

Figura A.5 – Processo de transcrição de proteína.



Fonte: Elaborada pela autora.

APÊNDICE B

Conceitos da Álgebra Abstrata e Códigos Corretores de Erros

A construção de códigos corretores de erros está fundamentada em estruturas algébricas portanto, as definições a seguir, extraídas de [69] e [106], são importantes para o entendimento desse processo. Este apêndice é destinado a introduzir os conceitos básicos de álgebra abstrata e álgebra linear que são importantes para o entendimento dos códigos BCH sobre corpos. Para tanto, inicialmente serão introduzidos os conceitos de grupo, anel e corpo e, em seguida, trata-se da teoria fundamental da álgebra linear para o entendimento dos códigos corretores de erros lineares.

Grupo

Considerando os conceitos da álgebra abstrata, uma estrutura algébrica consiste de modelos abstratos de conjuntos aos quais estão associados uma ou mais operações que satisfazem determinados axiomas.

***Definição* B.1** *Um* grupo *G é um conjunto munido com uma operação $* : G \times G \to G$ definida sobre os pares de elementos do conjunto, e satisfaz quatro propriedades:*

1. *Associatividade: $a * (b * c) = (a * b) * c$;*
2. *Identidade: O elemento identidade e satisfaz $a * e = e * a = a$ para todo a no conjunto;*
3. *Inverso: Se a pertence a G, então existe b em G tal que $a * b = b * a = e$.*

Se G tem um número finito de elementos, então ele é um grupo finito e o número de elementos em G é a ordem de G. O grupo é dito grupo abeliano *se satisfaz, adicionalmente, a propriedade da comutatividade, ou seja, $a * b = b * a$ para todo a e b no grupo.*

***Exemplo* B.1** *O* $(\mathbb{Z}_n, +)$ *é um grupo formado pelos números inteiros entre* 0 *e* $n-1$ *cuja soma é feita módulo* n*.*

***Exemplo* B.2** *O conjunto dos números inteiros sob a operação de adição é um grupo, entretanto, esse mesmo conjunto sob a operação de multiplicação não forma um grupo.*

Anéis

***Definição* B.2** *Um* anel A *é um conjunto com duas operações definidas*

$$\begin{array}{rccc} +: & A\times A & \to & A \\ & (a,b) & \mapsto & a+b \end{array} \qquad e \qquad \begin{array}{rccc} \cdot: & A\times A & \to & A \\ & (a,b) & \mapsto & a\cdot b \end{array}$$

chamadas, respectivamente, de adição e multiplicação, tal que satisfaz as seguintes propriedades:

1. *A é um grupo abeliano com relação à adição;*
2. *Fechado: Para qualquer a e b definidos em A, o produto ab também está em A;*
3. *Associativo: $a\cdot(b\cdot c) = (a\cdot b)\cdot c, \quad \forall\ a,b,c \in A$;*
4. *Distributivo: $a\cdot(b+c) = a\cdot b + a\cdot c$ ou $(b+c)\cdot a = b\cdot a + c\cdot a, \quad \forall\ a,b,c \in A$.*

Todo elemento em um anel tem um inverso na operação de adição porém, na multiplicação, um inverso é definido apenas em um anel com unidade. Se existe um elemento denotado por $1 \in A$ tal que $a\cdot 1 = 1\cdot a = a, \quad \forall\ a \in A$ diremos que A é um anel com unidade. Um elemento $a \in A$, será dito invertível se existir um elemento $b \in A$ tal que $a\cdot b = b\cdot a = 1$. E b é dito *inverso* de a.

Em um anel, a operação de adição é sempre comutativa, mas não necessariamente a operação de multiplicação também é. Um anel comutativo é aquele no qual a multiplicação é comutativa, ou seja, $a\cdot b = b\cdot a$ para todo a, b em A.

***Exemplo* B.3** *O conjunto dos números inteiros* $\mathbb{Z}$*, racionais* $\mathbb{Q}$*, reais* $\mathbb{R}$ *e complexos* $\mathbb{C}$ *munidos com as operações de adição e multiplicação usuais são exemplos de anéis. No entanto, o conjunto dos números naturais* $\mathbb{N}$ *munido com operações de adição e multiplicação dos inteiros não forma um anel pois não existem simétricos dos elementos, nem elemento neutro para a adição.*

Corpos

***Definição* B.3** *Um anel comutativo com unidade, em que todo elemento não nulo possui inverso multiplicativo é denominado um* corpo*. Define-se a ordem do corpo como sendo*

o número de elementos do corpo. Um corpo com um número finito de elementos será dito corpo finito.

***Definição* B.4** *Um* corpo $\mathbb{F}$ *é um conjunto com duas operações definidas: adição e multiplicação, tal que satisfaz as seguintes propriedades:*

1. $\mathbb{F}$ *é um grupo abeliano na adição;*
2. *O conjunto é fechado na multiplicação e os elementos não nulos são um grupo abeliano na multiplicação;*
3. *Distributivo:* $a \cdot (b + c) = a \cdot b + a \cdot c$ *ou* $(b + c) \cdot a = b \cdot a + c \cdot a$.

Um corpo com número finito de elementos q é chamando corpo finito ou corpo de Galois e é denotado por $GF(q)$ *ou* $\mathbb{F}_q$.

Em qualquer corpo finito, o número de elementos é uma potência de número primo. Assim, se q é um número primo e m é um positivo inteiro, o menor subcorpo de $\mathbb{F}_{q^m}$ é $\mathbb{F}_q$ e q é a característica de $\mathbb{F}_q$.

***Exemplo* B.4** *O conjunto dos inteiros* $\mathbb{Z}$ *(positivo, negativo e zero) forma um anel comutativo com identidade sob as operações usuais de adição e multiplicação contudo, não forma um corpo visto que nem todos elementos tem inverso multiplicativo.*

O menor corpo consiste de dois elementos e é denotado por $\mathbb{F}_2$. Esse corpo é constituído do elemento zero e do elemento um, cujas tabelas da operação de adição e multiplicação estão na Tabela B.1.

$+$	0	1
0	0	1
1	1	0

$\cdot$	0	1
0	0	0
1	0	1

Tabela B.1 – Tabela de operações $\mathbb{F}_2$.

Ao longo desta Tese a estrutura do corpo $\mathbb{F}_4$ foi amplamente utilizada. É importante notar que nessa estrutura as operações não são módulo 4 isto porque o $\mathbb{F}_2$ está contido em $\mathbb{F}_4$, ou seja, $\mathbb{F}_4$ é uma extensão do $\mathbb{F}_2$. As operações de adição e multiplicação entre os elementos do $\mathbb{F}_4$ estão na Tabela B.2.

$+$	0	1	α	α^2
0	0	1	α	α^2
1	1	0	α^2	α
α	α	α^2	0	1
α^2	α^2	α	1	0

$\cdot$	0	1	α	α^2
0	0	0	0	0
1	0	1	α	α^2
α	0	α	α^2	1
α^2	0	α^2	1	α

Tabela B.2 – Tabela de operações $\mathbb{F}_4$.

***Definição* B.5** *Seja $\alpha \in \mathbb{F}_q^*$, em que $\mathbb{F}_q^* = \mathbb{F}_q \setminus \{0\}$ é um corpo finito, define-se a* ordem *do elemento α como sendo o menor inteiro n tal que $\alpha^n = 1$.*

Considerando α um elemento não-nulo do corpo finito $\mathbb{F}_q$ com ordem n. Então, n divide $q - 1$.

***Definição* B.6** *Seja $\mathbb{F}$ um corpo, um subconjunto de $\mathbb{F}$ é chamado subcorpo se, com as operações de $\mathbb{F}$, também é um corpo. O corpo original $\mathbb{F}$ é então chamando de* corpo de extensão.

***Exemplo* B.5** *Considerando o corpo dos racionais $\mathbb{Q}$, reais $\mathbb{R}$ e complexos $\mathbb{C}$, têm-se que $\mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$. Neste caso, $\mathbb{C}$ é uma extensão de $\mathbb{R}$, $\mathbb{C}$ é uma extensão de $\mathbb{Q}$ e $\mathbb{R}$ é uma extensão de $\mathbb{Q}$.*

***Definição* B.7** *Um* elemento primitivo *de um corpo $\mathbb{F}_q$ é um elemento α tal que todo elemento do corpo exceto o zero pode ser expresso como uma potência de α. Portanto, as potências do elemento primitivo geram todos os elementos não-nulos de $\mathbb{F}_q$.*

Espaços Vetoriais

***Definição* B.8** *Seja um corpo K, cujos elementos são denominados de escalares, e um conjunto V, cujos elementos são denominados de vetores. Diz-se que V é um* espaço vetorial *sobre K, ou um K-*espaço vetorial*, se existir uma operação de adição em V,*

$$\begin{aligned} + : \ V \times V &\to V \\ (\mathbf{v}, \mathbf{w}) &\mapsto \mathbf{v} + \mathbf{w}, \end{aligned}$$

e uma operação de multiplicação de escalares por elementos de V, resultando em um vetor de V,

$$\begin{aligned} \cdot : \ K \times V &\to V \\ (\lambda, \mathbf{v}) &\mapsto \lambda \cdot \mathbf{v}, \end{aligned}$$

satisfazendo às seguintes propriedades:

1. *Associatividade da adição:* $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}), \quad \forall\ \mathbf{u}, \mathbf{v}, \mathbf{w} \in V$;
2. *Comutatividade da adição:* $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}, \quad \forall\ \mathbf{u}, \mathbf{v} \in V$;
3. *Existe um elemento neutro,* $0 \in V$, *tal que* $\mathbf{u} + 0 = \mathbf{u}, \quad \forall\ \mathbf{u} \in V$;
4. *Dado um elemento* $\mathbf{u} \in V$, *existe um elemento inverso* $-\mathbf{u}$, *chamado simétrico de* $\mathbf{u}$, *tal que,* $\mathbf{u} + (-\mathbf{u}) = 0$;

5. *Dados* $\lambda, \mu \in K$ *e* $\mathbf{u} \in V$*, vale:* $(\lambda + \mu) \cdot \mathbf{u} = \lambda \cdot \mathbf{u} + \mu \cdot \mathbf{u}$*;*

6. *Dados* $\lambda \in K$ *e* $\mathbf{u}, \mathbf{v} \in V$*, vale:* $\lambda \cdot (\mathbf{u} + \mathbf{v}) = \lambda \cdot \mathbf{u} + \lambda \cdot \mathbf{v}$*;*

7. *Dados* $\lambda, \mu \in K$ *e* $\mathbf{u} \in V$*, vale:* $(\lambda \cdot \mu) \cdot \mathbf{u} = \lambda \cdot (\mu \cdot \mathbf{u})$*;*

8. *Para todo* $\mathbf{u} \in V$*,* $1 \cdot \mathbf{u} = \mathbf{u}$*, em que* 1 *é a unidade de* K*.*

***Exemplo* B.6** *Os exemplos de espaços vetoriais mais comuns são os* $\mathbb{R}$*-espaços vetoriais* $\mathbb{R}^n$ *e os* $\mathbb{C}$*-espaços vetoriais* $\mathbb{C}^n$*. Esses são casos particulares de uma classe mais geral de espaços vetoriais, a dos* K*-espaços vetoriais* K^n*, em que* K *é um corpo arbitrário.*

***Definição* B.9** *Um subespaço vetorial de um* K*-espaço vetorial* V *é um subconjunto não vazio* W *de* V*, que, com as operações de adição e multiplicação por escalares de* V*, é também um* K*-espaço vetorial.*

Um subconjunto não vazio W de um espaço vetorial V é um subespaço vetorial se é satisfeita a seguinte condição,

$$\forall \mathbf{u}, \mathbf{v} \in W, \quad \forall \lambda \in K, \quad \mathbf{u} + \lambda \cdot \mathbf{v} \in W.$$

Seja V um K-espaço vetorial, dados $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$, dizemos que $\mathbf{v}_1, \ldots, \mathbf{v}_n$ são *linearmente independentes*, se for satisfeita a seguinte relação,

$$\lambda_1 \mathbf{v}_1 + \cdots + \lambda_n \mathbf{v}_n = \mathbf{0},$$

tal que $\lambda_1 = \cdots = \lambda_n = 0$ e $\lambda_1, \ldots, \lambda_n \in K$.

Anéis de Polinômios

***Definição* B.10** *Seja* $\mathbb{F}$ *um corpo e* x *uma indeterminada. Define-se o polinômio* $p(x)$ *com coeficientes em* $\mathbb{F}$ *na indeterminada* x *como*

$$p(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + \cdots + a_n x^n$$

em que $n \in \mathbb{Z}^+$ *e* $a_i \in \mathbb{F}, \quad \forall\ i = 0, 1, \ldots, n.$

Dados dois polinômios $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ e $q(x) = b_0 + b_1 x + \cdots + b_m x^m$, prova-se que $p(x) = q(x)$ se $a_i = b_i \ \ \forall\ i$. Seja $\mathbb{F}[x] = \{p(x); a_i \in \mathbb{F}; \quad \forall\ i = 1, \ldots, n\}$, ou seja, $\mathbb{F}[x]$ é o conjunto de todos os polinômios na indeterminada x com coeficientes em $\mathbb{F}$. Note Ao considerar $p(x)$ e $q(x)$ definidos anteriormente, define-se as seguintes operações:

$$\begin{aligned} p(x) + q(x) &= \sum_{i=0}^{\max\{n,m\}} (a_i + b_i) x^i \\ p(x) \cdot q(x) &= \sum_{i=0}^{n+m} c_i x^i \end{aligned}$$

em que $c_i = \sum_{j=0}^{i} a_j \cdot b_{i-j} x^i$. Assim, o conjunto $\mathbb{F}[x]$ munido com as operações definidas acima é um anel.

***Definição* B.11** *Um* polinômio primitivo $p(x)$ *sobre* $\mathbb{F}_q$ *é um polinômio primo com a seguinte propriedade: no corpo de extensão construído módulo* $p(x)$ *o elemento do corpo representado por* x *é um elemento primitivo.*

Elementos primitivos são úteis para a construção de corpos. A partir de um elemento primitivo é possível determinar os demais elementos do corpo que são potências do elemento primitivo.

***Exemplo* B.7** *A construção do corpo de extensão* $\mathbb{F}_{16}$ *a partir do* $\mathbb{F}_2$ *dar-se-à por meio de um polinômio primitivo de grau* 4 *como o* $p(x) = x^4 + x + 1$. *É possível perceber que esse polinômio não tem zero em* $\mathbb{F}_2$, $\mathbb{F}_4$ *e* $\mathbb{F}_8$ *mas, terá zero em* $\mathbb{F}_{2^4}$, *isto é, no corpo* $\mathbb{F}_{16}$ *existe algum elemento que satisfaz* $p(x) = 0$. *Os elementos do corpo de extensão tem ordem* 1, 3, 5 *ou* 15 *e aqueles com ordem* 15 *são primitivos.* α *é considerado o elemento primitivo e os elementos de* $\mathbb{F}_{16}$ *são:*

$$
\begin{array}{llll}
\alpha = \alpha & \alpha^9 = \alpha^3 + \alpha \\
\alpha^2 = \alpha^2 & \alpha^{10} = \alpha^2 + \alpha + 1 \\
\alpha^3 = \alpha^3 & \alpha^{11} = \alpha^3 + \alpha^2 + \alpha \\
\alpha^4 = \alpha + 1 & \alpha^{12} = \alpha^3 + \alpha^2 + \alpha + 1 \\
\alpha^5 = \alpha^2 + \alpha & \alpha^{13} = \alpha^3 + \alpha^2 + 1 \\
\alpha^6 = \alpha^3 + \alpha^2 & \alpha^{14} = \alpha^3 + 1 \\
\alpha^7 = \alpha^3 + \alpha + 1 & \alpha^{15} = 1 \\
\alpha^8 = \alpha^2 + 1 &
\end{array}
$$

O grupo dos elementos diferente de zero de $\mathbb{F}_q$ sobre a multiplicação é um grupo cíclico. A partir destes elementos é possível obter a seguinte fatoração:

$$x^{q-1} - 1 = (x - \alpha)(x - \alpha^2) \cdots (x - \alpha^{q-1}).$$

***Definição* B.12** *Seja* $\mathbb{F}_q$ *um corpo, e* $\mathbb{F}_Q$ *um corpo de extensão de* $\mathbb{F}_q$. *Se* β *está em* $\mathbb{F}_Q$, *o polinômio primo* $f(x)$ *de menor grau sobre* $\mathbb{F}_q$ *com* $f(\beta) = 0$ *é chamado* polinômio minimal *de* β *sobre* $\mathbb{F}_q$.

***Exemplo* B.8** *O polinômio minimal do número complexo i sobre o corpo dos reais é* $x^2 + 1$.

Códigos Corretores de Erros

Nos sistemas de comunicação, uma mensagem é transmitida por um canal que está sujeito a interferências, comumente chamadas de ruído. Os ruídos fazem com que a

mensagem recebida seja diferente da mensagem enviada. Portanto, para a transmissão da informação com confiabilidade, existe a necessidade de desenvolver métodos capazes de detectar e corrigir esses erros. Uma alternativa é a codificação para o controle de erros, que envolve o uso de um codificador de canal no transmissor e um algoritmo de decodificação no receptor.

Os códigos gerados pelo codificador de canal são chamados códigos corretores de erros e podem ser classificados basicamente em códigos de bloco e códigos convolucionais. Essa classificação é baseada na presença ou não de memória nos codificadores, assim, os códigos de bloco são ditos sem memória e os códigos convolucionais são ditos com memória, pois um determinado bit codificado depende de um ou mais bits de informação anteriores combinados linearmente. Em geral, a classe de códigos mais utilizada pertence à classe dos códigos de bloco lineares.

Seja $\mathbb{F}$ um corpo finito com q elementos. Em um codificador de bloco, a sequência de informação é segmentada em blocos de mensagens com k bits, denotados por $\mathbf{u} = (u_0, u_1, \ldots, u_{k-1})$, em que $u_i \in \mathbb{F}_q$, $i = 0, 1, \ldots, k-1$, assim teremos q^k possíveis mensagens. Cada mensagem $\mathbf{u}$ é transformada em uma palavra-código $\mathbf{v}$ com n bits. Os $n-k$ bits introduzidos na mensagem $\mathbf{u}$ são chamados bits de verificação de paridade que são a redundância utilizada para o decodificador identificar se houve erros durante a transmissão e, se possível, corrigi-los.

***Definição* B.13** *Um código de bloco de comprimento n e q^k palavras código é dito* código linear $\mathcal{C} \subset \mathbb{F}^n$, *denotado por (n, k), quando $\mathcal{C}$ é um subespaço vetorial de dimensão k de $\mathbb{F}_q^n$.*

Os parâmetros de um código C são agrupados na terna (n, k, d), em que n é o comprimento de bloco, k é a dimensão e d representa a distância mínima do código.

***Definição* B.14** *Seja $\mathbf{x} = (x_0,\ x_1,\ \ldots,\ x_{n-1}) \in V^n$, com V um espaço vetorial sobre $\mathbb{F}_q$. O* peso de Hamming *de $\mathbf{x}$ é definido como o número de símbolos diferentes de zero, ou seja,*

$$w(\mathbf{x}) := |\{i;\ x_i \neq 0\}|,$$

e o peso *de um código linear C é o inteiro:*

$$w(C) := \min\{w(\mathbf{x});\ \mathbf{x} \in C \setminus \{\mathbf{0}\}\}.$$

***Exemplo* B.9** *Seja $\mathbf{v} = (1\ 0\ 0\ 1\ 1\ 1\ 0)$, então o peso de Hamming de $\mathbf{v}$ é $w(\mathbf{x}) = 3$.*

***Definição* B.15** *Seja $\mathbf{u},\ \mathbf{v} \in V^n$. A* distância de Hamming *entre $\mathbf{u}$ e $\mathbf{v}$ denotada por $d(\mathbf{u},\ \mathbf{v})$ é definida como o número de coordenadas em que $\mathbf{u}$ e $\mathbf{v}$ diferem, isto é,*

$$d(\mathbf{u},\ \mathbf{v}) = |\{i;\ \mathbf{u}_i \neq \mathbf{v}_i,\ 0 \leq i \leq n-1\}|. \tag{B.1}$$

***Definição* B.16** *Dado um código $\mathcal{C} \in V^n$, então a* distância mínima *de $\mathcal{C}$ denotada por $d_{\min}$ é dada por:*

$$d_{\min} = \min\{d(\mathbf{u},\ \mathbf{v});\ \mathbf{u},\ \mathbf{v} \in \mathcal{C},\ \mathbf{u} \neq \mathbf{v}\}. \tag{B.2}$$

A distância de Hamming é uma métrica, também chamada de métrica de Hamming, portanto, valem as seguintes propriedades para $\mathbf{u},\ \mathbf{v},\ \mathbf{w} \in V^n$:

1. $d(\mathbf{u},\ \mathbf{v}) \geq 0$ e $d(\mathbf{u},\ \mathbf{v}) = 0 \iff \mathbf{u} = \mathbf{v}$;
2. $d(\mathbf{u},\ \mathbf{v}) = d(\mathbf{v},\ \mathbf{u})$;
3. Desigualdade triangular: $d(\mathbf{u},\ \mathbf{v}) \leq d(\mathbf{v},\ \mathbf{w}) + d(\mathbf{w},\ \mathbf{v})$.

Qualquer conjunto de vetores, que formam uma base ordenada para o subespaço, pode ser usado para formar uma matriz $k \times n$ chamada de matriz geradora do código e denotada por $\mathbf{G}$. As palavras código são resultantes de uma combinação linear das linhas de $\mathbf{G}$ e o conjunto de q^k palavras código é chamado de código linear.

***Exemplo* B.10** *Considerando um código linear binário, isto é, símbolos em $\mathbb{F}_2$, e seja a matriz geradora:*

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Todas as palavras código são: 00000, 00111, 01001, 01110, 10010, 10101, 11011 e 11100, portanto, este é um código (5,3,2).

Seja $\mathcal{C}$ um código linear com distância mínima d. Então, $\mathcal{C}$ pode corrigir até $t = \lfloor \frac{d-1}{2} \rfloor$ erros e detectar até $d-1$ erros. Os parâmetros $[n, k, d]$ de um código linear devem satisfazer à seguinte desigualdade:

$$d \leq n - k + 1. \tag{B.3}$$

Visto que um código linear C é um subespaço vetorial, então, o mesmo tem um complemento ortogonal, $C^{\perp}$, que é o conjunto de todos os vetores ortogonais a C. O complemento ortogonal também é um subespaço de $\mathbb{F}_q^n$ e, portanto, ele é um código.

***Definição* B.17** *Seja $C \subset \mathbb{F}_q^n$ um código linear, define-se*

$$C^{\perp} = \{\mathbf{v} \in \mathbb{F}_q^n;\ \langle \mathbf{v}, \mathbf{u} \rangle = 0 \ \ \forall\ \mathbf{u} \in C\},$$

em que, $\langle \mathbf{v}, \mathbf{u} \rangle$ é o produto interno de $\mathbf{u}$ e $\mathbf{v}$. O subespaço vetorial $C^{\perp}$ de $\mathbb{F}_q^n$, ortogonal a C, é também um código linear e é chamado de código dual *de C.*

Da teoria de espaço vetorial, o código dual $C^{\perp}$ tem dimensão $n - k$. Seja $\mathbf{H}$ uma matriz cujas linhas são vetores que formam a base de $C^{\perp}$. Então, uma n-úpla, $\mathbf{c}$, é uma

palavra-código em C se e somente se $\mathbf{c}$ for ortogonal a cada vetor de linha de $\mathbf{H}$. A matriz $\mathbf{H}$ é chamada de matriz de verificação de paridade do código C e,

$$\mathbf{cH}^T = 0.$$

Teorema B.1 *A matriz geradora do código C é a matriz de verificação de paridade para o código dual $C^\perp$.*

A matriz verificação de paridade de um código C pode ser obtida a partir da sua matriz geradora escrita na forma padrão, ou seja, sendo $\mathbf{G} = [I_k|A]$ em que I_k denota a matriz identidade de ordem k, têm-se que $\mathbf{H} = [-A^T|I_{n-k}]$.

***Exemplo* B.11** *Uma matriz de verificação de paridade para o código linear do Exemplo B.10 é dada por:*

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

Os códigos cíclicos formam uma subclasse da classe de códigos lineares. A estrutura destes códigos está diretamente relacionada à estrutura dos corpos de Galois, desta forma, os algoritmos de codificação e decodificação para estes códigos são computacionalmente eficientes.

***Definição* B.18** *Um código linear $\mathcal{C}$ de comprimento n é* cíclico *se para todo $\mathbf{c} = (c_0, c_1, \ldots, c_{n-2}, c_{n-1}) \in \mathcal{C}$, o vetor $\mathbf{c}^{(1)} = (c_{n-1}, c_0, c_1, \ldots, c_{n-2}) \in \mathcal{C}$.*

Nesse contexto, um deslocamento cíclico de uma palavra código em $\mathcal{C}$ resulta em uma outra palavra código pertencente a $\mathcal{C}$. Por exemplo, seja $\mathcal{C}_1 = \{0000, 1010, 0101, 1111\}$ e $\mathcal{C}_2 = \{0000, 1001, 0110, 1111\}$, têm-se que $\mathcal{C}_1$ é um código cíclico, porém, $\mathcal{C}_2$ não é um código cíclico.

No códigos cíclicos, os polinômios são descritos da seguinte forma:

Polinômio gerador:	$g(x)$	$\text{grau}(g(x)) = n - k$,
Polinômio de verificação de paridade:	$h(x)$	$\text{grau}(h(x)) = k$,
Mensagem:	$a(x)$	$\text{grau}(a(x)) = k - 1$,
Palavra código:	$c(x)$	$\text{grau}(g(x)) = n - 1$.

***Definição* B.19** *Um comprimento de bloco n da forma $n = q^m - 1$, em que q é uma potência de primo é denominado* comprimento de bloco primitivo *para um código sobre $\mathbb{F}_q$. Um código cíclico sobre $\mathbb{F}_q$ de comprimento de bloco primitivo é denominado um código* cíclico primitivo.

Portanto, códigos cíclicos primitivos sobre $\mathbb{F}_q$ podem ser encontrados através da fatoração de $x^{q^m-1} - 1$. Como visto na Seção anterior,

$$x^{q^m-1} - 1 = (x - \alpha)(x - \alpha^2) \cdots (x - \alpha^{q^m-1})$$

em que α é um elemento primitivo de $\mathbb{F}_{q^m}$. O mesmo polinômio pode ser fatorado sobre $\mathbb{F}_q$ como:

$$x^{q^m-1} - 1 = m_1(x)m_2(x)\cdots m_s(x)$$

em que os $m_i(x)$ são os polinômios minimais de cada elemento α de $\mathbb{F}_{q^m}$ e seus conjugados, isto é, α^q, α^{q^2}, α^{q^3}, ..., $\alpha^{q^{l-1}}$ em que l é o menor inteiro tal que $\alpha^{q^l} = \alpha$. Portanto, existem

$$\sum_{i=1}^{s-1} \begin{pmatrix} s \\ s-i \end{pmatrix} = \sum_{i=1}^{s-1} \frac{s!}{(s-i)!\ i!}$$

códigos cíclicos não triviais sobre $\mathbb{F}_q$, ou seja, todos os códigos de comprimento $n = q^m - 1$ com exceção do código universal e do código consistindo da palavra toda nula.

***Exemplo* B.12** *É possível construir códigos cíclicos de comprimento $n = 15 = 2^4 - 1$ a partir dos elementos da extensão de Galois $\mathbb{F}_{2^4}$ usando como polinômio primitivo $f(x) = x^4 + x + 1$. Portanto, seja o corpo descritos no Exemplo B.7, o polinômio minimal de cada elemento de $\mathbb{F}_{16}$ é:*

α^i	$m_i(x)$
0	x
1	$x - 1$
$\alpha, \alpha^2, \alpha^4, \alpha^8$	$x^4 + x + 1$
$\alpha^3, \alpha^6, \alpha^9, \alpha^{12}$	$x^4 + x^3 + x^2 + x + 1$
α^5, α^{10}	$x^2 + x + 1$
$\alpha^7, \alpha^{11}, \alpha^{13}, \alpha^{14}$	$x^4 + x^3 + 1$

Portanto,

$$x^{15} - 1 = (x+1)(x^4 + x + 1)(x^4 + x^3 + x^2 + x + 1)(x^2 + x + 1)(x^4 + x^3 + 1),$$

e existem 30 códigos cíclicos, não triviais, de comprimento $n = 15$.

APÊNDICE C

Quociente de Rayleigh

O quociente de Rayleigh para uma dada matriz Hermitiana $\mathbf{A} \in \mathbb{R}^{n \times n}$ e um vetor não nulo $\mathbf{x}$ é definido como,

$$\max_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{C.1}$$

Uma vez que o quociente de Rayleigh é invariante a escala, o problema pode ser reduzido à busca na esfera unitária, como segue:

$$\max_{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x}|| = 1} \mathbf{x}^T \mathbf{A} \mathbf{x}. \tag{C.2}$$

Para provar que o quociente de Rayleigh é maximizado quando $\mathbf{x}$ é o autovetor correspondente ao maior autovalor de $\mathbf{A}$, considere que a decomposição espectral de $\mathbf{A}$ é dada da seguinte maneira:

$$\mathbf{A} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T, \tag{C.3}$$

em que $\mathbf{Q} = [\mathbf{q}_1 \;\; \mathbf{q}_2 \;\; \ldots \;\; \mathbf{q}_n]$ é uma matriz quadrada $n \times n$ cuja i-ésima coluna corresponde ao autovetor $\mathbf{q}_i$, e $\mathbf{\Lambda} = \text{diag}(\lambda_1 \;\; \lambda_2 \;\; \ldots \;\; \lambda_n)$ é uma matriz diagonal cujos elementos da diagonal são os autovalores correspondentes e estão ordenados do maior para o menor. Então, para qualquer vetor unitário $\mathbf{x}$,

$$\begin{aligned} \mathbf{x}^T \mathbf{A} \mathbf{x} &= \mathbf{x}^T \left(\mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T \right) \mathbf{x} \\ &= \left(\mathbf{x}^T \mathbf{Q} \right) \mathbf{\Lambda} \left(\mathbf{Q}^T \mathbf{x} \right) \\ &= \mathbf{y}^T \mathbf{\Lambda} \mathbf{y}, \end{aligned} \tag{C.4}$$

em que $\mathbf{y} = \mathbf{Q}^T \mathbf{x}$ é também um vetor unitário pois,

$$\begin{aligned} ||\mathbf{y}||^2 &= \mathbf{y}^T \mathbf{y} \\ &= \left(\mathbf{Q}^T \mathbf{x} \right)^T \left(\mathbf{Q}^T \mathbf{x} \right) \\ &= \mathbf{x}^T \mathbf{Q} \mathbf{Q}^T \mathbf{x} \\ &= \mathbf{x}^T \mathbf{x} \\ &= 1. \end{aligned} \tag{C.5}$$

Assim, o problema inicial descrito na Equação (C.2) se reduz a,

$$\max_{\mathbf{y}\in\mathbb{R}^n:||\mathbf{y}||=1} \mathbf{y}^T\mathbf{\Lambda}\mathbf{y}. \tag{C.6}$$

Ou seja, considerando o vetor unitário $\mathbf{y} = [y_1 \quad y_2 \quad \dots \quad y_n]^T$,

$$\mathbf{y}^T\mathbf{\Lambda}\mathbf{y} = \sum_{i=1}^{n} \lambda_i y_i^2, \tag{C.7}$$

tal que

$$y_1^2 + y_2^2 + \cdots + y_n^2 = 1. \tag{C.8}$$

Como $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n$, a função objetivo da Equação (C.6) atinge seu máximo valor quando $\mathbf{y} = [1 \quad 0 \quad \dots \quad 0]$ ou $\mathbf{y} = [-1 \quad 0 \quad \dots \quad 0]$. Nesse caso, $\mathbf{y}^T\mathbf{\Lambda}\mathbf{y} = \lambda_1$. Assim,

$$\mathbf{x} = \mathbf{Q}\mathbf{y} = \pm\mathbf{q}_1. \tag{C.9}$$

Portanto, o quociente de Rayleigh $\mathbf{x}^T\mathbf{A}\mathbf{x}$ tem valor máximo igual a λ_1 (maior autovalor de $\mathbf{A}$) e ocorre quando $\mathbf{x} = \pm\mathbf{q}_1$ (maior autovetor de $\mathbf{A}$).

APÊNDICE D

Distância de Levenshtein

A distância de Levenshtein é uma métrica da semelhança entre duas palavras que podem ter comprimentos diferentes. Sendo assim, essa medida reflete o menor número de deleções, inserções e substituições necessárias para transformar uma palavra em outra. Aqui consideramos que o peso de uma deleção ou inserção é igual a um e a substituição tem peso dois, isto porque, um erro de substituição será interpretado como uma deleção seguida de uma inserção na mesma posição.

***Definição* D.1** *Considerando duas sequências* $\mathbf{u} \in \mathcal{A}^n$ *e* $\mathbf{v} \in \mathcal{A}^N$ *de comprimento* n *e* N *respectivamente, a* distância de Levenshtein *entre* $\mathbf{u}$ *e* $\mathbf{v}$ *é definida como,*

$$d_{\mathsf{L}}(\mathbf{u},\mathbf{v}) = d_{\mathbf{u},\mathbf{v}}(i,j) = \begin{cases} \max(i,j), & se\ \min(i,j) = 0 \\ \min \begin{cases} d_{\mathbf{u},\mathbf{v}}(i-1,j)+1 \\ d_{\mathbf{u},\mathbf{v}}(i,j-1)+1, \\ d_{\mathbf{u},\mathbf{v}}(i-1,j-1)+p(i,j) \end{cases} & c.c. \end{cases} \tag{D.1}$$

em que, $p(i,j) = \begin{cases} 0, & se\ u_i = v_j \\ 2, & c.c. \end{cases}$ *e* $d_{\mathbf{u},\mathbf{v}}(i,j)$ *é a distância entre os primeiros* i *caracteres de* $\mathbf{u}$ *e os primeiros* j *caracteres de* $\mathbf{v}$.

Quanto menor a distância, mais semelhantes são as duas sequências que estão sendo comparadas. A distância mínima de Levenshtein é a menor distância de Levenshtein entre todos os pares possíveis de vetores em um determinado conjunto. Uma abordagem para calcular a semelhança entre duas sequências seria gerar todos os alinhamentos possíveis e depois escolher o melhor. No entanto, a quantidade de alinhamentos entre duas sequências é exponencial, e essa abordagem resultaria em um algoritmo intoleravelmente lento [101].

***Exemplo* D.1** *Neste exemplo vamos ilustrar a aplicação da distância de Levenshtein para calcular a distância entre as sequências* $\mathbf{u} = SATURDAY$ *e* $\mathbf{v} = SUNDAY$. *Na Figura D.1 têm-se a a matriz completa resultante de todas as interações do algoritmo. Os comprimentos das sequências são 8 e 6, respectivamente, portanto, teremos uma matriz*

Figura D.1 – Matriz de dimensão 8×6 para calcular a distância de Levenshtein entre as sequências $\mathbf{u} = SATURDAY$ e $\mathbf{v} = SUNDAY$.

		S	A	T	U	R	D	A	Y
	0	1	2	3	4	5	6	7	8
S	1	**0**	**1**	**2**	3	4	5	6	7
U	2	1	2	3	**2**	3	4	5	6
N	3	2	3	4	3	**4**	5	6	7
D	4	3	4	5	4	5	**4**	5	6
A	5	4	3	4	5	6	5	**4**	5
Y	6	5	4	5	6	7	6	5	**4**

8×6. *A sequência de maior comprimento está ao longo do preenchimento horizontal e a de menor comprimento está ao longo do preenchimento vertical.*

A primeira linha e a primeira coluna são inicializadas com inteiros correspondentes ao comprimento das sequências. Isso ocorre porque se uma das sequências estiver vazia significa que todos os caracteres da outra sequência devem ser deletados, por exemplo, se $\mathbf{v} = 0$ *e* $i > j = 0$ *então a distância é* $d_{\mathsf{L}} = i$.

Cada elemento da matriz é rotulado por um par ordenado (k, l). *A principal observação é que podemos calcular o valor da entrada* $(k + 1, l + 1)$ *analisando qual o valor mínimo armazenado nas entradas anteriores, usando a seguinte regra:*

- *à entrada* $(k, j + 1)$ *adiciona-se* 1*: entende-se que seria necessário uma inserção;*
- *à entrada* $(k + 1, j)$ *adiciona-se* 1*: entende-se que seria necessário uma deleção;*
- *à entrada* (k, j) *adiciona-se* 2 *se os caracteres forem diferentes ou adiciona-se* 0 *se os caracteres forem iguais: verifica-se se seria necessário uma substituição.*

Por fim, o último elemento da última linha corresponde à distância de Levenshtein entre as duas sequências, nesse caso, $d_{\mathsf{L}} = 4$. *A partir desse elemento, o caminho com menores penalidades corresponde à escolha ótima para reescrever uma sequência como a outra. Porém, para fazer essa reconstrução seria necessário armazenar toda a matriz. Nesse caso, o caminho está em negrito e é interpretado da seguinte maneira: o* Y *mantém, o* A *mantém, o* D *mantém, o* R *substitui por* N*, o* U *mantém, o* T *deleta, o* A *deleta e o* S *mantém.*

APÊNDICE E

Algoritmo UPGMA

O método de grupo de pares não ponderados com média aritmética (UPGMA, do inglês: *Unweighted Pair Group Method with Arithmetic Mean*) é uma abordagem direta para construção de dendrogramas, ou árvores, a partir de matrizes de distâncias [107]. Isto é, o UPGMA é um método de clusterização hierárquico aglomerativo, no qual, o dendrograma é construído considerando que cada amostra representa um *cluster* que serão agrupados com base em uma medida de distância à medida que aumenta a hierarquia. Sendo assim, o dendrograma construído reflete a estrutura presente em uma matriz de semelhança de pares (ou uma matriz de dissimilaridade). Em cada etapa, os dois *clusters* mais próximos são combinados em um *cluster* de nível superior.

Para decidir quais *clusters* devem ser combinados é necessária uma medida de dissimilaridade entre esses conjuntos. No algoritmo UPGMA, a distância entre quaisquer dois *clusters* $\mathcal{A}$ e $\mathcal{B}$ é dada pela média das distâncias entre todos os elementos dos respectivos *clusters*, ou seja,

$$D(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}||\mathcal{B}|}\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} d(x,y), \tag{E.1}$$

em que $x \in \mathcal{A}$, $y \in \mathcal{B}$ e $d(x,y)$ é a distância entre os elementos x e y dada na matriz de distâncias. Portanto, na primeira etapa do algoritmo, são agrupados os dois *clusters* com a menor distância entre si, por exemplo $\mathcal{A}$ e $\mathcal{B}$. Esse agrupamento forma uma nova unidade taxonômica operacional $\mathcal{AB}$. Em seguida, uma nova matriz de distância é calculada incluindo o novo *cluster* $\mathcal{AB}$ em vez de considerá-los separadamente como $\mathcal{A}$ e $\mathcal{B}$. A distância atualizada entre $\mathcal{AB}$ e um novo cluster $\mathcal{X}$ é dado pela média proporcional de $D(\mathcal{A},\mathcal{X})$ e $D(\mathcal{B},\mathcal{X})$, ou seja,

$$D(\mathcal{A}\cup\mathcal{B},\mathcal{X}) = \frac{|\mathcal{A}|D(\mathcal{A},\mathcal{X}) + |\mathcal{B}|D(\mathcal{B},\mathcal{X})}{|\mathcal{A}| + |\mathcal{B}|}. \tag{E.2}$$

Esse processo é repetido iterativamente até que se obtenha um único *cluster*.

***Exemplo* E.1** *Considerando sete elementos* (a, b, c, d, e, f, g) *cujas distâncias entre si estão relacionada na seguinte matriz de distâncias a seguir. Inicialmente, cada elemento representa um cluster, portanto, utiliza-se notação caligráfica para indicar cada um deles na matriz de distâncias.*

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{E}$	$\mathcal{F}$
$\mathcal{B}$	19					
$\mathcal{C}$	27	31				
$\mathcal{D}$	8	18	26			
$\mathcal{E}$	33	36	41	31		
$\mathcal{F}$	18	1	32	17	35	
$\mathcal{G}$	13	13	29	14	28	12

Passo 1

O primeiro agrupamento é composto pelos dois clusters mais próximos, isto é: $\mathcal{B}$ *e* $\mathcal{F}$*. Calcula-se, então, a nova matriz de distâncias que será reduzida em tamanho por uma linha e uma coluna devido ao agrupamento de* $\mathcal{B}$ *e* $\mathcal{F}$*. Nesse caso, as novas distâncias são calculadas da seguinte maneira:*

$$
\begin{aligned}
D(\mathcal{B}\cup\mathcal{F},\mathcal{A}) &= (19+18)/2 = 18.5\\
D(\mathcal{B}\cup\mathcal{F},\mathcal{C}) &= (31+32)/2 = 31.5\\
D(\mathcal{B}\cup\mathcal{F},\mathcal{D}) &= (18+17)/2 = 17.5\\
D(\mathcal{B}\cup\mathcal{F},\mathcal{E}) &= (36+35)/2 = 35.5\\
D(\mathcal{B}\cup\mathcal{F},\mathcal{G}) &= (13+12)/2 = 12.5
\end{aligned}
$$

	$\mathcal{A}$	$\mathcal{BF}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{E}$
$\mathcal{BF}$	18.5				
$\mathcal{C}$	27	31.5			
$\mathcal{D}$	8	17.5	26		
$\mathcal{E}$	33	35.5	41	31	
$\mathcal{G}$	13	12.5	29	14	28

Passo 2

O novo agrupamento é composto pelos dois clusters mais próximos, isto é: $\mathcal{A}$ *e* $\mathcal{D}$*. Calcula-se, então, a nova matriz de distâncias.*

$$
\begin{aligned}
D(\mathcal{A}\cup\mathcal{D},\mathcal{BF}) &= (18.5+17.5)/2 = 18\\
D(\mathcal{A}\cup\mathcal{D},\mathcal{C}) &= (27+26)/2 = 26.5\\
D(\mathcal{A}\cup\mathcal{D},\mathcal{E}) &= (33+31)/2 = 32\\
D(\mathcal{A}\cup\mathcal{D},\mathcal{G}) &= (13+14)/2 = 13.5
\end{aligned}
$$

	$\mathcal{AD}$	$\mathcal{BF}$	$\mathcal{C}$	$\mathcal{E}$
$\mathcal{BF}$	18			
$\mathcal{C}$	26.5	31.5		
$\mathcal{E}$	32	35.5	41	
$\mathcal{G}$	13.5	12.5	29	28

Passo 3

O novo agrupamento é composto pelos dois clusters mais próximos, isto é: $\mathcal{BF}$ *e* $\mathcal{G}$*. Calcula-se, então, a nova matriz de distâncias.*

$$
\begin{aligned}
D(\mathcal{BF} \cup \mathcal{G}, \mathcal{AD}) &= (2 \times 18 + 13.5)/3 = 16.5 \\
D(\mathcal{BF} \cup \mathcal{G}, \mathcal{C}) &= (2 \times 31.5 + 29)/3 = 30.67 \\
D(\mathcal{BF} \cup \mathcal{G}, \mathcal{E}) &= (2 \times 35.5 + 28)/3 = 33
\end{aligned}
$$

	$\mathcal{AD}$	$\mathcal{BFG}$	$\mathcal{C}$
$\mathcal{BFG}$	16.5		
$\mathcal{C}$	26.5	30.67	
$\mathcal{E}$	32	33	41

Passo 4

O novo agrupamento é composto pelos dois clusters mais próximos, isto é: $\mathcal{AD}$ *e* $\mathcal{BFG}$*. Calcula-se, então, a nova matriz de distâncias.*

$$
\begin{aligned}
D(\mathcal{AD} \cup \mathcal{BFG}, \mathcal{C}) &= 29 \\
D(\mathcal{AD} \cup \mathcal{BFG}, \mathcal{E}) &= 32.6
\end{aligned}
$$

	$\mathcal{ADBFG}$	$\mathcal{C}$
$\mathcal{C}$	29	
$\mathcal{E}$	32.6	41

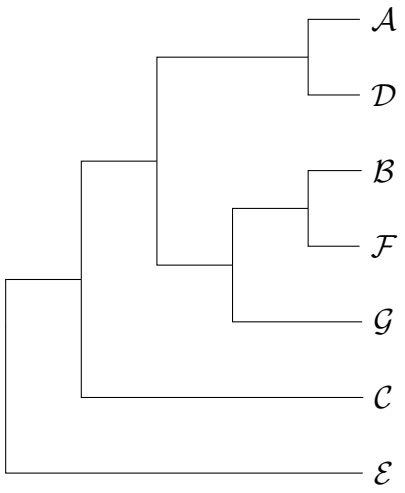
Passo 5

O novo agrupamento é composto pelos dois clusters mais próximos, isto é: $\mathcal{ADBFG}$ *e* $\mathcal{C}$*. Calcula-se, então, a nova matriz de distâncias.*

$$D(\mathcal{ADBFG} \cup \mathcal{C}, \mathcal{E}) = 34$$

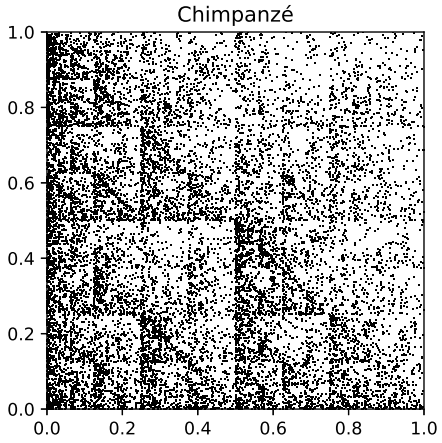
	$\mathcal{ADBFGC}$
$\mathcal{E}$	34

Por fim, o dendrograma que representa esse conjunto de sete elementos é dado por:

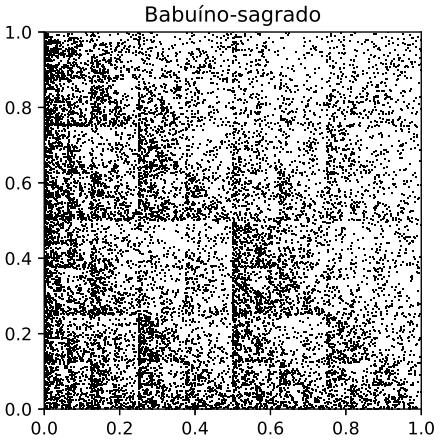


APÊNDICE F

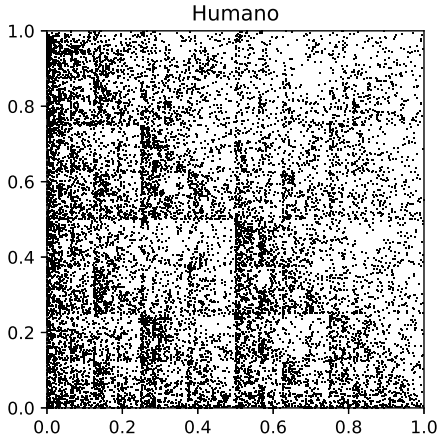
CGR de Sequências de DNA de Mamíferos



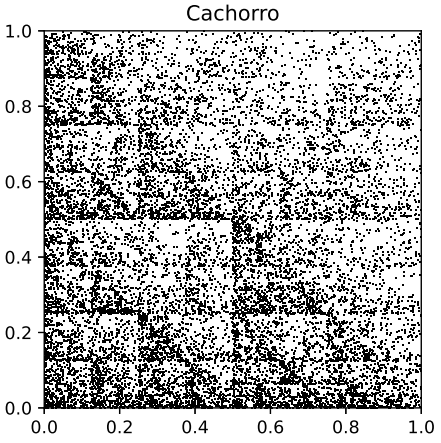
(a)



(b)

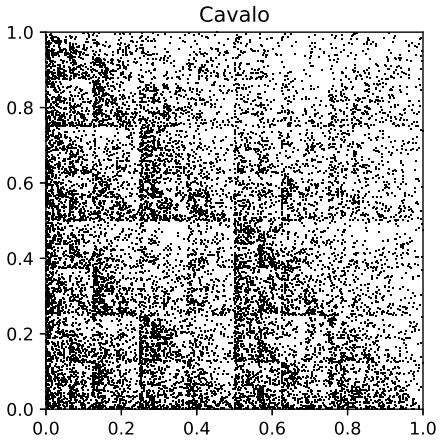


(c)

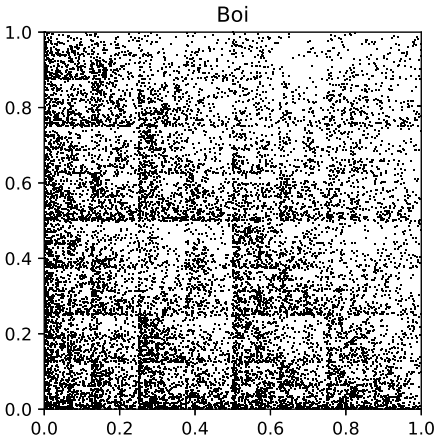


(d)

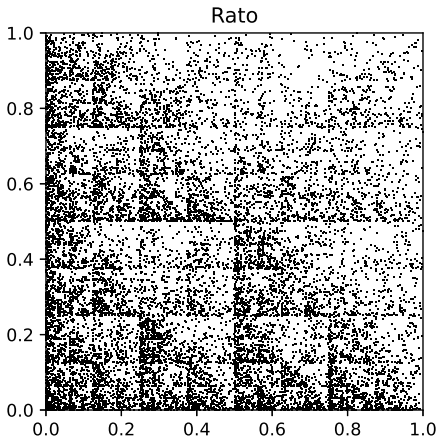
Fonte: Elaborada pela autora



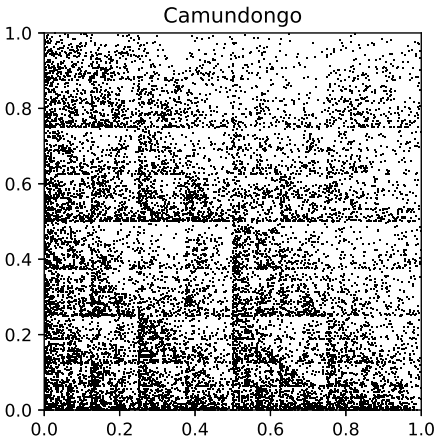
(e)



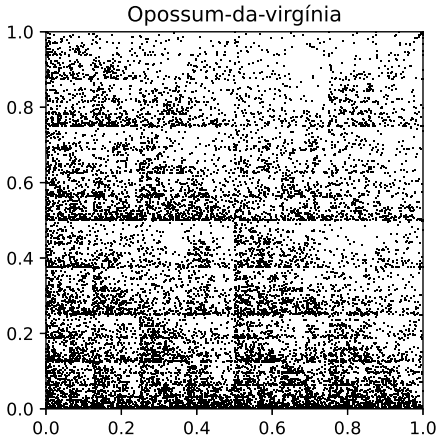
(f)



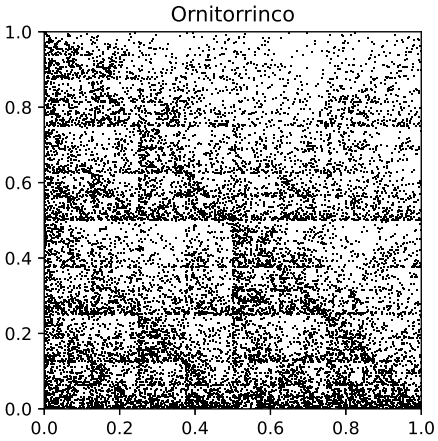
(g)



(h)



(i)



(j)

Fonte: Elaborada pela autora